Human Activity Recognition Using Conditional Random Fields and Privileged Information

DOCTORAL THESIS

submitted to

the designated by the General Assembly Composition of the

Department of Computer Science & Engineering Inquiry Committee

by

Michalis Vrigkas

in partial fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

February 2016

Αναγνώριση Ανθρώπινης Δραστηριότητας με Υπό Συνθήκη Τυχαία Πεδία και Προνομιακή Πληροφορία

Η ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Μηχανικών Η/Υ και Πληροφορικής Εξεταστική Επιτροπή

από τον

Μιχαήλ Βρίγκα

ως μέρος των Υποχρεώσεων για τη λήψη του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Φεβρουάριος 2016

Committees

Advisory Committee

1. **Christophoros Nikou**, Associate Professor, Department of Computer Science and Engineering, University of Ioannina, Greece (*supervisor*)
2. **Ioannis A. Kakadiaris**, Professor, Department of Computer Science, University of Houston, Houston, TX, USA
3. **Lisimachos-Paul Kondi**, Associate Professor, Department of Computer Science and Engineering, University of Ioannina, Greece

Examination Committee

1. **Christophoros Nikou**, Associate Professor, Department of Computer Science and Engineering, University of Ioannina, Greece (*supervisor*)
2. **Ioannis A. Kakadiaris,** Professor, Department of Computer Science, University of Houston, Houston, TX, USA
3. **Lisimachos-Paul Kondi**, Associate Professor, Department of Computer Science and Engineering, University of Ioannina, Greece
4. **Aristidis Likas**, Professor, Department of Computer Science and Engineering, University of Ioannina, Greece
5. **Konstantinos Blekas**, Associate Professor, Department of Computer Science and Engineering, University of Ioannina, Greece
6. **Antonis Argyros**, Professor, Department of Computer Science, University of Crete, Greece
7. **George Bebis**, Professor, Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA

Dedication

To my family,
for without their support, and love none of these would have happened.
To new coming experiences.

Στην οικογένειά μου,
χωρίς την δική τους υποστήριξη, και αγάπη τίποτα από αυτά δεν θα είχε συμβεί.
Στα καινούρια που ακολουθούν.

Acknowledgements

Walking through the path of knowledge has always been an interesting challenge for me. Despite the ups and downs, the difficulties and disappointments, the happy moments and bad ones, my journey has finally come to its end.

Completing my PhD thesis would have been impossible without the help and support of my supervisor, Prof. Christophoros Nikou, whom I would like to thank for teaching me how to think. I will always be grateful to him for his guidance and patience in transforming me into a scientist. His excellent professional ethic motivated me to vigorously engage with research in the Computer Vision field.

I would also like to thank Prof. Ioannis Kakadiaris, for his valuable advises, and help during my graduate studies. I was honored to collaborate with him and to learn from his experience. I would also like to thank Prof. Lisimachos-Paul Kondi for his advise and the exceptional collaboration during all these years. He has perfectly assisted me to all of my research needs. I would also like to thank Profs. Aristidis Likas, Konstandinos Blekas, Antonis Argyros, and George Bebis for serving in the examination committee of my dissertation.

Special thanks to all of my friends and lab-mates for being so supportive to me and for sharing some of the happiest days of our lives. All these countless hours of talking, brainstorming, and coffee/activity time sharing, have given me the strength I needed to "recharge my batteries" after a long and tiring day.

My last and most important words of love and gratitude are kept for my parents Stefanos and Angeliki, my sisters Giota and Katerina, my brothers in law George and George, my niece Adamandia and my life-partner Eleni. It is thanks to them I was able to come to the finish line of my PhD. Thank You!

Michalis Vrigkas,
Ioannina, February 2016

Contents

1 Introduction **1**
- 1.1 Human Activity Recognition from Video Sequences . . . 1
- 1.2 Thesis Contribution . . . 2

2 Background and Related Work in Action Recognition **7**
- 2.1 Introduction . . . 7
- 2.2 Human Activity Categorization . . . 9
- 2.3 Unimodal Methods . . . 10
  - 2.3.1 Space-Time Methods . . . 10
  - 2.3.2 Stochastic Methods . . . 15
  - 2.3.3 Rule-Based Methods . . . 18
  - 2.3.4 Shape-Based Methods . . . 20
- 2.4 Multimodal Methods . . . 23
  - 2.4.1 Affective Methods . . . 24
  - 2.4.2 Behavioral Methods . . . 25
  - 2.4.3 Methods Based on Social Networking . . . 27
  - 2.4.4 Multimodal Feature Fusion . . . 31
- 2.5 Discussion . . . 32

3 Matching Mixtures of Trajectories for Human Action Recognition **37**
- 3.1 Introduction . . . 37
- 3.2 Action Representation and Recognition . . . 38
  - 3.2.1 Motion Representation . . . 39
  - 3.2.2 Extraction of Motion Curves . . . 40
  - 3.2.3 Motion Curves Clustering . . . 40
  - 3.2.4 Matching of Motion Curves . . . 42
  - 3.2.5 Canonical Time Warping . . . 43
  - 3.2.6 Dimensionality Reduction . . . 44
- 3.3 Experimental Results . . . 44
  - 3.3.1 Evaluation over the Weizmann Dataset . . . 44
  - 3.3.2 Evaluation over the KTH Dataset . . . 47
  - 3.3.3 Evaluation over the UCF Sports Dataset . . . 52
  - 3.3.4 Evaluation over the UCF YouTube Dataset . . . 53

3.3.5 Parameter Estimation . . . 56
3.3.6 Discussion . . . 58
3.4 Conclusion . . . 60

4 Classifying Behavioral Attributes Using Conditional Random Fields 63
4.1 Introduction . . . 63
4.2 Behavior Recognition Using Conditional Random Fields . . . 64
4.2.1 Learning . . . 65
4.2.2 Inference . . . 66
4.3 Experimental Results . . . 67
4.3.1 Political Behavior Dataset . . . 67
4.3.2 Results and Discussion . . . 69
4.4 Conclusion . . . 70

5 Identifying Human Behaviors Using Hidden Conditional Random Fields 71
5.1 Introduction . . . 71
5.2 Behavior Recognition Using Hidden Conditional Random Fields . . . 72
5.2.1 Multimodal Hidden Conditional Random Fields . . . 73
5.2.2 Parameter Learning and Inference . . . 74
5.2.3 Multimodal Feature Extraction . . . 75
5.2.4 Audio-Visual Synchronization and Fusion . . . 76
5.3 Experimental Results . . . 78
5.3.1 Datasets . . . 78
5.3.2 Implementation details . . . 79
5.3.3 Model Selection . . . 80
5.3.4 Feature Pruning . . . 82
5.3.5 Comparison of Learning Frameworks . . . 86
5.4 Conclusion . . . 89

6 Human Activity Recognition Using Robust Adaptive Privileged Probabilistic Learning 93
6.1 Introduction . . . 93
6.2 Robust Privileged Probabilistic Learning . . . 95
6.2.1 HCRF+ Model Formulation . . . 96
6.2.2 Maximum Likelihood Learning . . . 98
6.2.3 Maximum Margin Learning . . . 98
6.2.4 Estimation of Regularization Parameters . . . 99
6.2.5 Inference . . . 100
6.2.6 Mapping of Discrete Features to Continuous Space . . . 101
6.3 Experimental Results . . . 102
6.3.1 Datasets . . . 102
6.3.2 Feature Selection . . . 103

6.3.3 Model Selection . . . 104
6.3.4 Evaluation of Privileged Information . . . 105
6.3.5 Comparison of Learning Frameworks . . . 107
6.4 Conclusion . . . 117

7 Active Privileged Learning of Human Activities from Weakly Labeled Samples 119
7.1 Introduction . . . 119
7.2 Active Privileged Leaning . . . 120
7.2.1 a-HCRF+ Model Formulation . . . 120
7.2.2 Learning and Inference . . . 122
7.2.3 Active Learning . . . 123
7.3 Experimental Results . . . 123
7.3.1 Datasets . . . 124
7.3.2 Implementation Details . . . 124
7.3.3 Results and Discussion . . . 125
7.4 Conclusion . . . 127

8 Exploiting Privileged Information for Facial Expression Recognition 129
8.1 Introduction . . . 129
8.2 Leaning to Transfer Privileged Information . . . 131
8.2.1 t-CRF+ Model Formulation . . . 131
8.2.2 Parameter Learning and Inference . . . 132
8.3 Experimental Results . . . 134
8.3.1 Datasets . . . 134
8.3.2 Baseline Approaches . . . 135
8.3.3 Model Selection . . . 135
8.3.4 Results and Discussion . . . 136
8.4 Conclusion . . . 139

9 Conclusions and Future Work 141
9.1 Conclusions . . . 141
9.2 Limitations and Directions for Future Work . . . 143

Appendices 145

A Conditional Distribution of the Privileged Information 145
A.1 Conditional Student's t-Distribution . . . 145

B Learning Using Privileged Information 147
B.1 SVM+ Formulation . . . 147

Bibliography 149

List of Figures

1.1 Decomposition of human activities. 3

2.1 Proposed hierarchical categorization of human activity recognition methods. 9

3.1 Overview of our approach. 39
3.2 Depiction of the LCSS matching between two motions considering that they should be within $\delta = 64$ time steps in the horizontal axis and their amplitudes should differ at most by $\varepsilon = 0.086$. 43
3.3 Sample frames from video sequences of the Weizmann dataset [1]. 46
3.4 Confusion matrices of the classification results for the Weizmann dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion. 47
3.5 The recognition accuracy with respect to the number of Gaussian components for the Weizmann dataset. 48
3.6 Sample frames from video sequences of the KTH dataset [2]. 49
3.7 Confusion matrices of the classification results for the KTH dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion. 50
3.8 The recognition accuracy with respect to the number of Gaussian components for the KTH dataset. 51
3.9 Sample frames from video sequences of the UCF Sports dataset [3]. 52
3.10 Confusion matrices of the classification results for the UCF Sports dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion. 53
3.11 The recognition accuracy with respect to the number of Gaussian components for the UCF Sports dataset. 54
3.12 Sample frames from video sequences of the UCF YouTube action dataset [4]. 56

3.13 Confusion matrices of the classification results for the UCF YouTube dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion. 57
3.14 The recognition accuracy with respect to the number of Gaussian components for the UCF YouTube dataset. 58
3.15 Average percentage of matched curves for TMAR(LCSS) and TMAR(CTW), when the BIC criterion is used, for (a) Weizman, (b) KTH, (c) UCF Sports and (d) UCF YouTube datasets, respectively. 60
3.16 Execution times per action in seconds for TMAR(LCSS), TMAR(CTW) and TMAR(PCA), when the BIC criterion is used, for (a) KTH, (b) UCF Sports and (c) UCF YouTube datasets, respectively. 61

4.1 Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral 64
4.2 Graphical representation of the model. The observed features are represented by $\mathbf{x}$ and the unknown labels are represented by $\mathbf{y}$. Temporal edges exist also between the labels and the observed features across frames. . . . 66
4.3 Tree-like graphical representation of the model. The observed features are represented by $\mathbf{x}$ and the unknown labels are represented by $\mathbf{y}$. 66
4.4 Sample frames from the proposed *Parliament* dataset. (Top row) Friendly, (middle row) Aggressive, and (bottom row) Neutral. 67
4.5 Distribution of classes *friendly*, *aggressive*, and *neutral*. (a) Bhattacharyya distance between classes for all video samples. (b) Distribution of each class against the others (bottom row) after projection onto a common subspace using PCA. The main diagonal shows how data are distributed within each class. 68
4.6 Confusion matrices of the classification results for the CRF model employing (a) only unary potentials, (b) only unary potentials without spatio-temporal features, (c) the full model without spatio-temporal pairwise features, and (d) the full model. 70

5.1 Graphical representation of the chain structure model. The grey nodes are the observed features and the unknown labels represented by x and y, respectively. The white nodes are the unobserved hidden variables h. . . . 73
5.2 Representative examples of feature pruning. (a) The original features and (b) the pruned features for the *Parliament* dataset [5] (top row) and the TV human interaction dataset [6] (bottom row). Feature pruning may reduce the number of features by 29% on average. 77
5.3 Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral. 79

5.4 Sample frames from the TVHI dataset. (a) Hand shake, (b) High five, (c) Hug, and (d) Kiss. 79
5.5 Comparison of the per class number of visual features before and after pruning for (a) the *Parliament* and (b) the TVHI datasets. 80
5.6 Synchronization offsets between audio and video features for some sample video sequences of the *Parliament* (top row) and TVHI (bottom row) datasets. The circle indicates a delay of (a) -44 frames, (b) +13 frames, (c) −13 frames and (d) +37 frames. 81
5.7 Canonical variates of audio and visual features for two sample videos of (a) the *Parliament* and (b) the TVHI (bottom row) datasets. Notice the high correlation between audio and visual features obtained by the projection. . 82
5.8 Confusion matrices for the classification results of the proposed SAVAR approach for the *Parliament* dataset [5], after feature pruning, using 5-fold cross validation. 88
5.9 Confusion matrices for the classification results of the proposed SAVAR approach for the *Parliament* dataset [5], after feature pruning, using LOSO cross validation. 89
5.10 Confusion matrices for the classification results of the proposed SAVAR approach for the TVHI dataset [6], after feature pruning. 90

6.1 Robust learning using privileged information. Given a set of training examples and a set of additional information about the training samples (left) our system can successfully recognize the class label of the underlying activity without having access to the additional information during testing (right). We explore three different forms of privileged information (e.g., audio signals, human poses, and attributes) by modeling them with a Student's t-distribution and incorporating them into the HCRF+ model. . . . 94
6.2 Graphical representation of the chain structure model. The grey nodes are the observed features (x_i), the privileged information (x_i^*), and the unknown labels (y), respectively. The white nodes are the unobserved hidden variables (h). 96
6.3 Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral. 102
6.4 Sample frames from the TVHI dataset. (a) Hand shake, (b) High five, (c) Hug, and (d) Kiss. 102
6.5 Sample frames from the TPI dataset. (a) Approach, (b) Depart, (c) Kick, (d) Push, (e) Shake hands, (f) Hug, (g) Exchange objects, and (h) Punch. . 103
6.6 Sample frames from the USAA dataset. (a) Birthday party, (b) Graduation party, (c) Music performance, (d) Non-music performance, (e) Parade, (f) Wedding ceremony, (g) Wedding dance, and (h) Wedding reception. 104

6.7 Comparison of the recognition accuracy of the four different variants of the proposed method and standard HCRF model with respect to the number of hidden states for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7], and (d) the USAA [8] datasets. The text in parentheses in the legend of each figure corresponds to the type of information used both for training and testing. 106
6.8 Recognition performance of the proposed maximum likelihood variant as function of the regularization parameter and the number of hidden states for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7] and (d) the USAA [8] datasets. 107
6.9 Recognition performance of the proposed max-margin variant as function of the regularization parameter and the number of hidden states for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7] and (d) the USAA [8] datasets. 108
6.10 Confusion matrices for the classification results of the proposed HCRF+ approach for the *Parliament* dataset [5] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants. . . 110
6.11 Confusion matrices for the classification results of the proposed HCRF+ approach for the TVHI dataset [6] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants. 112
6.12 Confusion matrices for the classification results of the proposed HCRF+ approach for the TPI dataset [7] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants. 113
6.13 Confusion matrices for the classification results of the proposed HCRF+ approach for the USAA dataset [8] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants. 114

7.1 Graphical representation of the chain structure model. The grey nodes are the observed features (x_i and x_i^*), and the unknown labels (y). The white nodes are the hidden variables (h). 121
7.2 Comparison of classification accuracies with respect to the number of unlabeled data for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7], and (d) the USAA [8] datasets. 126
7.3 Confusion matrices for the classification results for the best split of the proposed a-HCRF+ model for the Parliament [5] (first row), the TVHI [6] (second row), the TPI [7] (third row), and the USAA [8] (fourth row) datasets. Right column corresponds to a-HCRF+ (entropy) and left column corresponds to a-HCRF+ (ratioCP) variants, respectively. 128

8.1 An overview of the proposed framework. 130

8.2 Graphical representation of the chain structure CRF model. The grey nodes are the observed features (x_i) and the white nodes are unknown labels (y_i), respectively. 131
8.3 Proposed t-CRF+ model. First, a standard chain structure CRF model is trained on the privileged feature space ($\mathcal{X}^*$) with parameters $\mathbf{w_p}$. Then, the privileged knowledge is transferred to the original feature space ($\mathcal{X}$). The square nodes correspond to the unary and pairwise potentials, which are conditioned on their hyper-parameters $\mathbf{w_p}$ and $\mathbf{w_o}$, respectively. 133
8.4 Illustration of ROC curves for (a) AVEC 2011 [9] and (b) CK+ [10] datasets. 138
8.5 Comparison of recognition performance accuracies (%) of each class for (a) AVEC 2011 [9] and (b) CK+ [10] datasets. 139

List of Tables

2.1 Summary of previous surveys. 9
2.2 Comparison of unimodal methods. 33
2.3 Comparison of multimodal methods. 34
2.4 Human activity recognition datasets. 35

3.1 Recognition accuracy over the Weizmann dataset. 46
3.2 p-values for measuring the statistical significance of the proposed methods for the Weizmann dataset. The null hypothesis appears in the first column of the table. 48
3.3 Statistical measurements of the recognition results for each of the proposed approaches for the Weizmann dataset. 48
3.4 Recognition results over the KTH dataset. 49
3.5 p-values for measuring the statistical significance of the proposed methods for the KTH dataset. 51
3.6 Statistical measurements of the recognition results for each of the proposed approaches for the KTH dataset. All values are expressed in percentages. . 52
3.7 Recognition results over the UCF Sport dataset. 54
3.8 p-values for measuring the statistical significance of the proposed methods for the UCF Sports dataset. 55
3.9 Statistical measurements of the recognition results for each of the proposed approaches for the UCF Sports dataset. All values are expressed in percentages. 55
3.10 Recognition results over the UCF YouTube dataset. 55
3.11 p-values for measuring the statistical significance of the proposed methods for the UCF YouTube dataset. 56
3.12 Statistical measurements of the recognition results for each of the proposed approaches for the UCF YouTube dataset. All values are expressed in percentages. 56
3.13 Parameters δ and ε for the KTH dataset estimated using cross validation. . 58
3.14 Parameters δ and ε for the UCF Sports dataset estimated using cross validation. 59
3.15 Parameters δ and ε for the UCF Youtube dataset estimated using cross validation. 59

4.1 Behavior classification accuracies (%) using the graphical model with only temporal edges (4.2) and the full graphical model (4.1). 69
4.2 Comparison between variants of the proposed method. 69

5.1 Types of audio and visual features used for human behavior recognition. The numbers in parentheses indicate the dimension of the features. 76
5.2 Recognition accuracy of the proposed HCRF model with respect to the number of hidden states (h=$\{3 \dots 10\}$) for the *Parliament* dataset [5] using 5-fold and LOSO cross validation, before feature pruning and after feature pruning. 84
5.3 Recognition accuracy of the proposed HCRF model with respect to the number of hidden states (h=$\{4 \dots 10\}$) the TVHI dataset [6] before feature pruning and after feature pruning. 85
5.4 Classification results on the *Parliament* dataset [5]. 86
5.5 Classification results on the TVHI dataset [6]. 87
5.6 p-values of the proposed method for the *Parliament* dataset [5]. 88
5.7 p-values of the proposed method for the TVHI dataset [6]. 89

6.1 Types of features used for human activity recognition for each dataset. The numbers in parentheses indicate the dimension of the features. The checkmark corresponds to the usage of the specific information as regular or privileged. Privileged features are used only during training. 105
6.2 Comparison of the classification accuracies (%) on Parliament dataset [5]. . 109
6.3 Comparison of the classification accuracies (%) on TVHI dataset [6]. 111
6.4 Comparison of the classification accuracies (%) on TPI dataset [7]. 112
6.5 Comparison of the classification accuracies (%) on USAA dataset [8]. . . . 113
6.6 p-values of the proposed method for the *Parliament* dataset [5]. 115
6.7 p-values of the proposed method for the TVHI dataset [6]. 116
6.8 p-values of the proposed method for the TPI dataset [7]. 116
6.9 p-values of the proposed method for the USAA dataset [8]. 117

7.1 Comparison of the classification accuracies (%) for the Parliament [5], TVHI [6], TPI [7], and USAA [8] datasets. The results were averaged for all different configurations (mean ± standard deviation). 127

8.1 Comparison of feature combinations for classifying facial expressions and affective states on AVEC 2011 [9], and CK+ [10] datasets. The crossmark indicates the absence of privileged information during training. 136
8.2 Comparison of the classification accuracies and the area under the ROC curve (%) for the AVEC 2011 [9] and the CK+ [10] datasets. 137
8.3 p-values of the proposed method for the AVEC 2011 [9] and the CK+ [10] datasets. 138

List of Algorithms

1 Action learning 45
2 Action categorization 45
3 Feature pruning 76
4 Audio-visual synchronization 78
5 Pool-based active learning using a-HCRF+ 124
6 Transferring knowledge from $\mathcal{X}^*$ to $\mathcal{X}$ using t-CRF+ 134

Abstract

Michalis S. Vrigkas.
PhD, Department of Computer Science & Engineering, University of Ioannina, Greece. February, 2016.
Thesis Title: Human activity recognition using conditional random fields and privileged information.
Thesis Supervisor: Christophoros Nikou.

Recognizing human activities from video sequences or still images is a challenging task due to problems such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system.

In the first part of this thesis, after a review of the state-of-the-art methods, a learning-based framework for action representation and recognition relying on time series of optical flow motion features is presented. In the learning step, the motion curves representing each action are clustered using Gaussian mixture modeling (GMM). In the recognition step, the optical flow curves of a probe sequence are also clustered using a GMM, then each probe sequence is projected onto the training space and the probe curves are matched to the learned curves using a non-metric similarity function based on the longest common subsequence, which is robust to noise and provides an intuitive notion of similarity between curves.

Next, a human behavior recognition method with an application to political speech videos is presented. The behavior of a subject is modeled using a conditional random field (CRF). To evaluate the performance of the model, a novel behavior dataset is introduced, which includes low resolution video sequences depicting different people speaking in the Greek parliament. The subjects of the *Parliament* dataset are labeled as friendly, aggressive or neutral depending on the intensity of their political speech.

An extension of the aforementioned human behavior recognition method using multimodal features is also presented. Individual and social behaviors of a subject are modeled using a hidden conditional random field (HCRF). Each video is represented by a vector of spatio-temporal visual features along with audio features. To remove irrelevant features a feature pruning method based on the spatio-temporal neighborhood of each feature in a video sequence is presented. The proposed framework assumes that human movements are highly correlated with sound emissions and canonical correlation analysis is employed

to find relationship between the audio and video features prior to fusion.

Besides the classical learning frameworks, a novel method based on the learning using privileged information (LUPI) paradigm for recognizing complex human activities is proposed that handles missing information during testing. A supervised probabilistic approach that integrates LUPI into an HCRF model is presented. The proposed model employs a self-training technique for automatic estimation of the regularization parameters of the objective function. Moreover, the method provides robustness to outliers by modeling the conditional distribution of the privileged information by a Student's t-density function. Different forms of additional information were investigated.

In many human activity recognition systems the size of the unlabeled training data may be significantly large due to expensive human effort required for data annotation. Moreover, the insufficient data collection process from heterogenous sources may cause dissimilarities between training and testing data. To address these limitations, a novel probabilistic approach that combines LUPI and active learning is proposed. A pool-based privileged active learning approach is presented for semi-supervising learning of human activities from multimodal labeled and unlabeled data.

In the last part of this dissertation, the LUPI paradigm is also investigated for solving biometric applications such as facial expression recognition. As facial image sequences may contain information for heterogeneous sources, facial data may be asymmetrically distributed between training and testing, as it may be difficult to maintain the same quality and quantity of information. To this end, a novel probabilistic classification method that combined the LUPI framework and conditional random fields is proposed to indirectly propagate knowledge from privileged to regular feature space. Each feature space owns specific parameter settings, which are combined together through a Gaussian prior, to train the proposed t-CRF+ model and allow the different tasks to share parameters and improve classification performance.

Εκτεταμενη Περιληψη στα Ελληνικα

Μιχάλης Βρίγκας του Στεφάνου και της Αγγελικής.
PhD, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Φεβρουάριος, 2016.
Τίτλος Διατριβής: Αναγνώριση ανθρώπινης δραστηριότητας με υπό συνθήκη τυχαία πεδία και προνομιακή πληροφορία.
Επιβλέπων Καθηγητής: Χριστόφορος Νίκου.

Το πρόβλημα της αναγνώρισης και του εντοπισμού της ανθρώπινης δραστηριότητας από εικονοσειρές και απλές εικόνες, είναι μία δύσκολη διαδικασία, λόγω προβλημάτων όπως ύπαρξη θορύβου στα δεδομένα, αλλαγές στην κλίμακα, την φωτεινότητα και την εμφάνιση. Πολλές εφαρμογές παρακολούθησης εικονοσειρών, αλληλεπίδρασης ανθρώπου-υπολογιστή και διάφορα ρομποτικά συστήματα, απαιτούν αλγορίθμους για την αναγνώριση της ανθρώπινης δραστηριότητας.

Στο πρώτο μέρος της διατριβής, και ύστερα από μια λεπτομερή και διεξοδική ανάλυση των μεθοδολογιών αναγνώρισης της ανθρώπινης δραστηριότητας, περιγράφεται μια μέθοδος βασισμένη στην σύγκριση τροχιών για αναγνώριση της ανθρώπινης δραστηριότητας. Η μέθοδος βασίζεται στην περιγραφή μιας ανθρώπινης δράσης από χρονοσειρές βασισμένες στην οπτική ροή. Αρχικά, στο βήμα εκπαίδευσης, οι καμπύλες κίνησης που αναπαριστούν μια δράση ομαδοποιούνται από μία μικτή κανονική κατανομή. Στη φάση της αναγνώρισης, οι καμπύλες κίνησης μιας καινούριας εικονοσειράς ομαδοποιούνται επίσης με μία μικτή κανονική κατανομή και το υπό κατηγοριοποίηση μοντέλο συγκρίνεται με όλα τα μοντέλα της βάσης εκπαίδευσης χρησιμοποιώντας ένα μέτρο ομοιότητας που βασίζεται στη μεγαλύτερη κοινή υπακολουθία μεταξύ των μέσων καμπυλών των μικτών κατανομών.

Στη συνέχεια, παρουσιάζεται μια μέθοδος για την αναγνώριση της ανθρώπινης συμπεριφοράς σε πολιτικές ομιλίες. Η συμπεριφορά ενός ατόμου μοντελοποιείται χρησιμοποιώντας υπό συνθήκη τυχαία πεδία.Για την αξιολόγηση της απόδοσης του μοντέλου, δημιουργήθηκε ένα καινούριο σύνολο δεδομένων το οποίο αποτελείται από ομιλίες βουλευτών της ελληνικής βουλής και σε κάθε υποκείμενο ανατίθεται μία από τρεις κατηγορίες συμπεριφοράς, όπως φιλικός, επιθετικός και ουδέτερος.

Έπειτα, παρουσιάζεται μια επέκταση της προαναφερθείσας μεθόδου, χρησιμοποιώντας δεδομένα από πολλαπλές πηγές. Η συμπεριφορά ενός ατόμου αναπαριστάται εισάγοντας ένα επίπεδο κρυμμένων καταστάσεων για την μοντελοποίηση των κρυμμένων δυναμικών του προβλήματος της αναγνώρισης. Επίσης, κάθε εικονοσειρά αναπαριστάται ταυτόχρονα

με οπτικά χαρακτηριστικά και με χαρακτηριστικά ήχου. Για την απομάκρυνση περιττών χαρακτηριστικών που εμφανίζονται κυρίως λόγω θορύβου σε κάθε εικονοπλαίσιο, προτείνεται μια μέθοδος για την μείωση του αριθμού τους βασισμένη στην χωρική και χρονική γειτνίαση των σημείων αυτών. Για τον αυτόματο συγχρονισμό και την συγχώνευση των οπτικών και ηχητικών σημάτων χρησιμοποιήθηκε η μέθοδος της ανάλυσης κανονικής συσχέτισης.

Τα περισσότερα μοντέλα ταξινόμησης δεν λαμβάνουν υπόψη τους την ανισορροπία που υπάρχει στην δομή των δεδομένων για εκπαίδευση και έλεγχο. Για το λόγο αυτό, προτείνεται ένα μοντέλο, το οποίο χρησιμοποιεί επιπλέον δεδομένα (προνομιακή πληροφορία) μόνο στην φάση της εκπαίδευσης, ενώ στην φάση του ελέγχου αυτή η πληροφορία δεν είναι διαθέσιμη. Η προτεινόμενη μέθοδος είναι βασισμένη στην εκπαίδευση με χρήση προνομιακής πληροφορίας και είναι ανθεκτική σε δεδομένα τα οποία δεν ακολουθούν το κυρίαρχο μοντέλο, όπως θόρυβος ή ελλιπή δεδομένα, μοντελοποιώντας την υπό συνθήκη κατανομή της προνομιακής πληροφορίας χρησιμοποιώντας την κατανομή *Student's-t*. Η συγκεκριμένη προσέγγιση είναι γενική και δεν περιορίζεται στην χρήση μόνο ενός είδους προνομιακής πληροφορίας. Επίσης, προτείνεται μία μέθοδος για αυτόματη εκτίμηση της τιμής των παραμέτρων ομαλοποίησης μέσα από μια διαδικασία αυτοεκπαίδευσης από το σύνολο δεδομένων.

Σε πολλά συστήματα αναγνώρισης ανθρώπινης δραστηριότητας το μέγεθος των μη επισημασμένων δεδομένων εκπαίδευσης μπορεί να είναι σημαντικά μεγάλο, κυρίως λόγω της επίπονης και χρονοβόρας ανθρώπινης προσπάθειας για την περιγραφή των δεδομένων. Η ανεπαρκής, σε πολλές περιπτώσεις, διαδικασία συλλογής δεδομένων από ετερογενείς πηγές μπορεί να προκαλέσει ανομοιότητες μεταξύ των δεδομένων εκπαίδευσης και ελέγχου. Για την αντιμετώπιση αυτών των περιορισμών, προτείνεται μια νέα προσέγγιση, η οποία συνδυάζει τη μάθηση με τη χρήση προνομιακής πληροφορίας και την ενεργή μάθηση για την αναγνώριση ανθρώπινων δραστηριοτήτων από πολυτροπικά και μη χαρακτηρισμένα με κάποια ετικέτα δεδομένα.

Στο τελευταίο μέρος της διατριβής, χρησιμοποιείται η προνομιακή πληροφορία για την επίλυση βιομετρικών εφαρμογών, όπως η αναγνώριση εκφράσεων του προπώπου. Καθώς οι εικόνες προσώπων μπορεί να περιέχουν ετερογενείς πληροφορίες, τα δεδομένα του προσώπου μπορεί να είναι ανομοιόμορφα κατανεμημένα μεταξύ της φάσης εκπαίδευσης και του ελέγχου, και έτσι μπορεί να είναι είναι δύσκολο να διατηρηθεί η ίδια ποιότητα και ποσότητα των πληροφοριών. Για το λόγο αυτό, προτείνεται μια μέθοδος ταξινόμησης, η οποία συνδιάζει την προνομιακή πληροφορία και τα υπό συνθήκη τυχαία πεδία, για να να διαδώσει έμμεσα γνώση από τον προνομιακό στον αρχικό χώρο των δεδομένων. Κάθε χώρος έχει συγκεκριμένες παραμέτρους, οι οποίες συνδέονται μεταξύ τους μέσω μιας Γκαουσιανής κατανομής, για να επιτρέψουν στις διαφορετικές διαδικασίες μάθησης να μοιραστούν τις διαφορετικές παραμέτρους και να βελτιώθει η ταξινόμηση.

Chapter 1

Introduction

1.1 Human Activity Recognition from Video Sequences

1.2 Thesis Contribution

1.1 Human Activity Recognition from Video Sequences

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations. Because it provides information about the identity, personality, and psychological state, it is difficult to extract. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning. As a result of this research, many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system.

Among various classification techniques two main questions arise: "What action?" (i.e., the recognition problem) and "Where in the video?" (i.e., the localization problem). When attempting to recognize human activities one must determine the kinetic states of a person, so that the computer can efficiently recognize this activity. Human activities such as "walking" or "running" arise very naturally in daily life and are relatively easy to recognize. On the other hand, more complex activities such as "peeling an apple" are more difficult to identify. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event [11].

Most of the work in human activity recognition assumes a figure-centric scene of uncluttered background, where the actor is free to perform an activity. The development of a fully automated human activity recognition system, capable of classifying a person's activities with low error, is a challenging task due to problems such as background clutter,

partial occlusion, changes in scale, viewpoint, lighting and appearance, and frame resolution. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. Moreover, intra- and inter-class similarities make the problem amply challenging. That is, actions within the same class may be expressed by different people with different body movements and actions between different classes may be difficult to distinguish as they may be represented by similar information. The way that humans perform an activity depends on their habits, and this makes the problem of identifying the underlying activity quite difficult to determine. Also, the construction of a visual model for learning and analyzing human movements in real time with inadequate benchmark datasets for evaluation are challenging tasks.

To overcome these problems a task is required that consists of three components, namely: (i) background subtraction [12, 13], in which the system attempts to separate the parts of the image that are invariant over time (background) from the objects that are moving or changing (foreground); (ii) human tracking, in which the system locates human motion over time [14, 15, 16]; and (iii) human action and object detection [17, 18, 19], in which the system is able to localize a human activity in an image.

The goal of human activity recognition is to examine activities from video sequences or still images. Motivated by this fact, human activity recognition systems aim to correctly classify input data into its underlying activity category. Depending on their complexity, human activities are categorized into: (i) gestures; (ii) atomic actions; (iii) human-to-object or human-to-human interactions; (iv) group actions; (v) behaviors; and (vi) events. Figure 1.1 visualizes the decomposition of human activities according to their complexity.

Gestures are considered as primitive movements of the body parts of a person that may correspond to a particular action of this person [20]. Atomic actions are movements of a person describing a certain motion that may be part of more complex activities [21]. Human-to-object or human-to-human interactions are human activities that involve two or more persons or objects [6]. Group actions are activities performed by a group or persons [22]. Human behaviors refer to physical actions that are associated with the emotions, personality and psychological state of the individual [23]. Finally, events are high level activities that describe social actions between individuals and indicate the intention or the social role of a person [24].

1.2 Thesis Contribution

This thesis focuses on the development of efficient human activity recognition methods using several graphical probabilistic models such as Gaussian mixture models (GMM) [25], conditional random fields (CRF) [26], and hidden conditional random fields (HCRF) [27]. Conditional random fields are more suitable to encode sequential human activities by representing the dependencies between the observations and the actual class label with a structured graphical model. They are formed as a collection of different feature functions and are able to work well with complex features that may be extracted form different

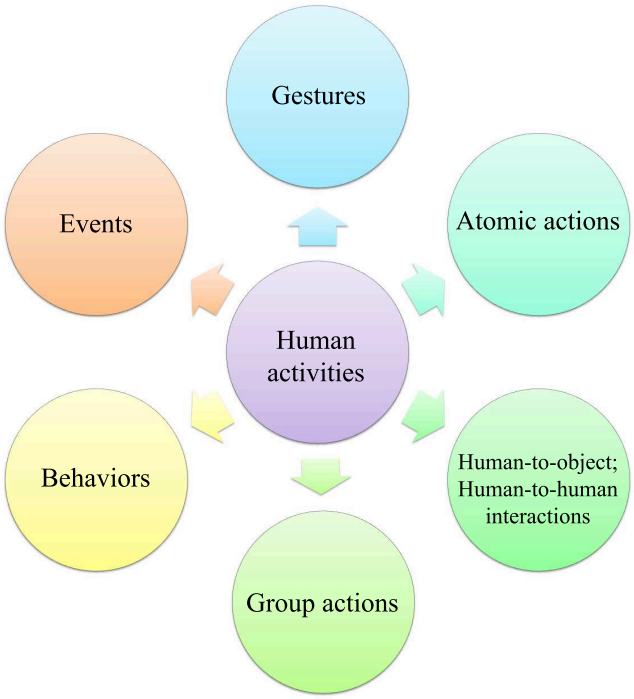


Figure 1.1: Decomposition of human activities.

sources. A key issue when modeling human activities with such models is which features are more informative for this task. In this dissertation, a set of diverse and multimodal features are used and innovative classification tasks are combined together to address the problems in classification of activities due to dissimilarities in training and testing features that may occur from the data acquisition procedure.

For a better insight of the existing methodologies, in Chapter 2, we overview the related work for human activity recognition and present a new taxonomy of the related methods. This helps to categorize those human actions that play a significant role in evaluating human activity recognition systems. Moreover, we present a complete list of human activity datasets categorizing them according to the kind of activities they may represent. We also set the bases for an ideal human activity classification system and discuss the strengths and weaknesses of each category separately.

In Chapter 3, human activities are represented by a set of clustered motion curves. These motion curves consist of time series of optical flow motion features, which are grouped using Gaussian mixture modeling to represent the different activities. To avoid flaws in the representation of human motion due to the sensitivity of optical flow analysis to noise and partial occlusions, a non-metric similarity function based on the longest common subsequence is used. The learned motion curves are matched to a new probe sequence by detecting similar pairs of curve segments. The advantage of the propose method is that it is able to handle motion curves of different lengths and is robust to outliers. Since a human actions may non uniformly occur within a video sequence, the continuity of the curves along time is ensured by tracking the optical flow features. More-

over, canonical time warping is employed for spatio-temporal alignment of the motion curves and dimensionality reduction is applied to remove outliers from the motion curves and reduce their lengths.

In Chapter 4, a human behavior recognition method with application to political speeches is presented and a novel behavior dataset, called the *Parliament* dataset is also introduced. This dataset is a collection of low resolution video sequences depicting different people speaking in the Greek parliament. Each sequence was manually labeled with one of three behavioral labels, which correspond to friendly, aggressive and neutral states. The discrimination between friendly and aggressive labels is not straightforward in political speeches as the subjects perform similar movements in both cases. To model the underlying human behavior, a fully connected conditional random field (CRF) is employed, where different labels for each video frame were considered. This makes the model more suitable to handle video sequences with more than one label per video sequence.

In Chapter 5, a multimodal human action recognition method based on hidden conditional random fields (HCRF) is proposed. To reduce the number of irrelevant features that may occur during data acquisition constraints, a feature selection technique based on the spatio-temporal neighborhood of the visual features is employed. Moreover, the correlation of audio and visual information is investigated for recognizing complex human activities. However, due to the different frame rate that each modality may have, audio and video features are temporally aligned such that the correlation between sound emissions and human movements is maximized. The combination of both visual and audio information reinforces the intuition that human behaviors are more easily identified as they are characterized by complex actions of movements and sound emissions.

In Chapter 6, a novel classification model that exploits learning using privileged information (LUPI) is introduced. Within this framework, training data are enhanced with additional information also called privileged information, that may reflect on natural or auxiliary properties about classes and members of the classes of the training data. Privileged information is available only during training but never during testing. This kind of learning is of high importance as it resembles the human ability of learning by exploiting only useful information, which are provided by a strong teacher during training, while only a few information may be available during testing. The LUPI framework is used in a probabilistic manner by incorporating it in a hidden conditional random field model, while maximum likelihood and maximum margin approaches are employed to learn the model's parameters. The regularization parameters are automatically inferred through a self-training procedure directly form the training data.

In Chapter 7, a semi-supervising method using active learning and privileged information for recognizing human activities is presented. Knowing the ground truth labeling for all training examples in advance may not always be feasible for large scale heterogeneous and unconstrained data. Moreover, to reduce tedious human effort in data annotation, which may be time consuming and computationally expensive, a combination of learning using privileged information and active learning is proposed. The benefit of the proposed

methodology is twofold. First, it is able to cope with information that is not available during testing and second, it addresses the problem of missing labels during training. Both procedures are performed simultaneously through a unified semi-supervising pool-based active learning technique.

In Chapter 8, a novel classification method for indirect propagation of knowledge from privileged to regular feature space is introduced with application to biometrics such as facial expression and affective recognition. Each domain is treated separately and the learned privileged weights are used to penalize the original feature space through a Gaussian prior. Thus, samples that have a good evidence to distinguish between classes for both privileged and original feature spaces contribute heavier to the learning process, while samples that are harder to separate have less effect to the leaning model. The proposed method is not limited to the use of any specific form of auxiliary information.

Finally, Chapter 9 summarizes this thesis, provides some possible extensions of the proposed methodologies and overviews the directions for future work.

Chapter 2

Background and Related Work in Action Recognition

2.1 Introduction

2.2 Human Activity Categorization

2.3 Unimodal Methods

2.4 Multimodal Methods

2.5 Discussion

2.1 Introduction

There are several surveys in the human activity recognition literature. Gavrila [28] separated the research in 2D (with and without explicit shape models) and 3D approaches. Aggarwal and Cai [29], presented a new taxonomy focusing on human motion analysis, tracking from single view and multi view cameras and recognition of human activities. Similar in spirit to the previous taxonomy, Wang *et al.* [30] proposed a hierarchical action categorization hierarchy. The survey of Moeslund *et al.* [31] mainly focused on pose-based action recognition methods and proposed a four-fold taxonomy including initialization of human motion, tracking, pose estimation, and recognition methods.

A fine separation between the meanings of "action" and "activity" was proposed by Turaga *et al.* [32], where the activity recognition methods were categorized according to their degree of activity complexity. Poppe [33] characterized human activity recognition methods into two main categories, describing them as "top-down" and "bottom-up". On the other hand, Aggarwal and Ryoo [34] presented a tree structured taxonomy, where the human activity recognition methods were categorized into two big subcategories, the

"single layer" approaches and the "hierarchical" approaches, each of which have several layers of categorization.

Modeling 3D data is also a new trend and it was extensively studied by Chen *et al.* [35] and Ye *et al.* [36]. As the human body consists of limbs connected with joints, one can model these parts using stronger features, which are obtained from depth cameras, and create a 3D representation of the human body, which is more informative than the analysis of 2D activities carried out in the image plane. Aggarwal and Xia [37] recently presented a categorization of human activity recognition methods from 3D stereo and motion capture systems with the main focus on methods that exploit 3D depth data. To this end, Microsoft Kinect has played a significant role in motion capture of articulated body skeletons using depth sensors.

Although much research has been focused on human activity recognition systems from video sequences, human activity recognition from static images remains an open and very challenging task. Most of the studies of human activity recognition are associated with facial expression recognition and/or pose estimation techniques. Guo and Lai [38] summarized all the methods for human activity recognition from still images and categorized them into two big categories according to the level of abstraction and the type of features each method uses.

Jaimes and Sebe [39] proposed a survey for multimodal human computer interaction focusing on affective interaction methods from poses, facial expressions and speech. Pantic and Rothkrantz [40] performed a complete study in human affective state recognition methods that incorporate nonverbal multimodal cues such as facial and vocal expressions. Pantic *et al.* [41] studied several state-of-the-art methods of human behavior recognition including affective and social cues, and covered many open computational problems and how they can be efficiently incorporated into a human-computer interaction system. Zeng *et al.* [42] presented a review of state-of-the-art affective recognition methods that use visual and audio cues for recognizing spontaneous affective states and provided a list of related datasets for human affective expression recognition. Bousmalis *et al.* [43] proposed an analysis of non-verbal multimodal (i.e., visual and auditory cues) behavior recognition methods and datasets for spontaneous agreements and disagreements. Such social attributes may play an important role in analyzing social behaviors, which are the key to social engagement. Finally, a thorough analysis of the ontologies for human behavior recognition from the viewpoint of data and knowledge representation was presented by Rodríguez *et al.* [44].

Table 2.1 summarizes the previous surveys on human activity and behavior recognition methods sorted by chronological order. Most of these reviews summarize human activity recognition methods, without providing the strengths and the weaknesses of each category in a concise and informative way. Our goal is not only to present a new classification for the human activity recognition methods, but also to compare different state-of-the-art studies and understand the advantages and disadvantages of each method.

Table 2.1: Summary of previous surveys.

Authors	Year	Area of interest
Aggarwal and Cai [29]	1999	Human motion analysis and tracking from single and multi view data.
Gavrila [28]	1999	Shape model analysis from 2D and 3D data.
Pantic and Rothkrantz [40]	2003	Multimodal human affective state recognition.
Wang *et al.* [30]	2003	Human detection, tracking and activity recognition.
Moeslund *et al.* [31]	2006	Motion initialization, tracking, pose estimation, and recognition.
Pantic *et al.* [41]	2006	Investigation of affective and social behaviors for human-computer interactions.
Jaimes and Sebe [39]	2007	Multimodal affective interaction analysis for human-computer interactions.
Turaga *et al.* [32]	2008	Categorization of actions and activities according to their complexity.
Zeng *et al.* [42]	2009	Audio-visual affective recognition analysis.
Poppe [33]	2010	Action classification according to global or local representation of data.
Aggarwal and Ryoo [34]	2011	Gestures, human activities, actions and interactions analysis.
Bousmalis *et al.* [43]	2013	Audio-visual behavior analysis of spontaneous agreements and disagreements.
Chen *et al.* [35]	2013	Human body part motion analysis from depth image data.
Ye *et al.* [36]	2013	Human activity analysis from skeletal poses using depth data.
Aggarwal and Xia [37]	2014	Human activity analysis from stereo, motion capture, and depth sensors 3D data.
Guo and Lai [38]	2014	Understanding human activities from still images.
Rodríguez *et al.* [44]	2014	Representation of human behavior ontologies from knowledge-based techniques.

2.2 Human Activity Categorization

The human activity categorization problem has remained a challenging task in computer vision for more than two decades. Previous works on characterizing human behavior have shown great potential in this area. First, we categorize the human activity recognition methods into two main categories: (i) *unimodal* and (ii) *multimodal* activity recognition methods according to the nature of sensor data they employ. Then, each of these two categories is further analyzed into sub-categories depending on how they model human activities. Thus, we propose a hierarchical classification of the human activity recognition methods, which is depicted in Figure 2.1.

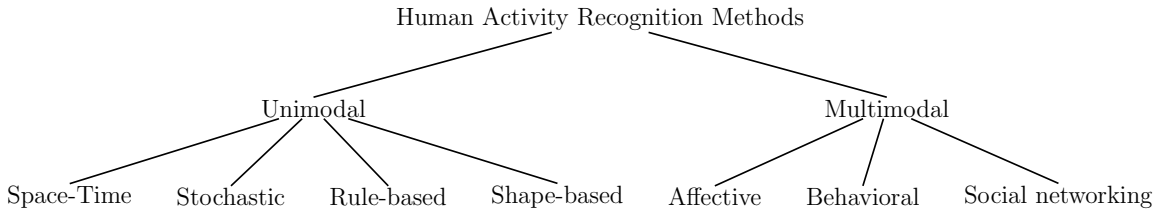


Figure 2.1: Proposed hierarchical categorization of human activity recognition methods.

Unimodal methods represent human activities from data of a single modality, such as images, and they are further categorized as: (i) *space-time*, (ii) *stochastic*, (iii) *rule-based*, and (iv) *shape-based methods*.

Space-time methods involve activity recognition methods, which represent human activities as a set of spatio-temporal features [45, 46] or trajectories [47, 48]. Stochastic methods recognize activities by applying statistical models to represent human actions (e.g., hidden Markov models) [49, 50]. Rule-based methods use a set of rules to describe

human activities [51, 52]. Shape-based methods efficiently represent activities with high-level reasoning by modeling the motion of human body parts [53, 54].

Multimodal methods combine features collected from different sources [55] and are classified into three categories: (i) *affective*, (ii) *behavioral*, and (iii) *social networking methods*.

Affective methods represent human activities according to emotional communications and the affective state of a person [23, 56]. Behavioral methods aim to recognize behavioral attributes, non-verbal multimodal cues such as gestures, facial expressions and auditory cues [5, 57]. Finally, social networking methods model the characteristics and the behavior of humans in several layers of human-to-human interactions in social events from gestures, body motion, and speech [6, 58].

Usually, the terms "activity" and "behavior" are used interchangeably in the literature [57, 59]. In this survey, we differentiate between these two terms in the sense that the term "activity" is used to describe a sequence of actions that correspond to specific body motion. On the other hand, the term "behavior" is used to characterize both activities and events that are associated with gestures, emotional states, facial expressions and auditory cues of a single person.

2.3 Unimodal Methods

Unimodal human activity recognition methods identify human activities from data of one modality. Most of the existing approaches represent human activities as a set of visual features extracted from video sequences or still images and recognize the underlying activity label using several classification models [60, 61]. Unimodal approaches are appropriate for recognizing human activities based on motion features. However, the ability to recognize the underlying class only from motion is on its own a challenging task. The main problem is how we can ensure the continuity of the motion along time as an action occurs uniformly or non-uniformly within a video sequence. Some approaches use snippets of motion trajectories [62, 63], while others use the full length of motion curves by tracking the optical flow features [64].

We classify unimodal methods into four broad categories: (i) *space-time*, (ii) *stochastic*, (iii) *rule-based*, and (iv) *shape-based approaches*. Each of these subcategories describes specific attributes of human activity recognition methods according to the type of representation each method uses.

2.3.1 Space-Time Methods

Space-time approaches focus on recognizing activities based on space-time features or on trajectory matching. They consider an activity in the 3D space-time volume, consisting of concatenation of 2D spaces in time. An activity is represented by a set of space-time features or trajectories extracted from a video sequence.

A plethora of human activity recognition methods based on space-time representation have been proposed in the literature [2, 65, 66, 67, 68]. A major family of methods relies on optical flow, which has proven to be an important cue. Efros *et al.* [65] recognized human actions from low-resolution sports video sequences using the nearest neighbor classifier, where humans are represented by windows of height of 30 pixels. The approach of Fathi and Mori [66] was based on mid-level motion features, which are also constructed directly from optical flow features. Moreover, Wang and Mori [69] employed motion features as input to hidden conditional random fields (HCRF) [27] and support vector machine (SVM) classifiers [25]. Real time classification and prediction of future actions was proposed by Morris and Trivedi [70], where an activity vocabulary is learned through a three-step procedure. Other optical flow-based methods which gained popularity were presented in [71, 72, 73]. An invariant in translation and scaling descriptor was introduced by Oikonomopoulos *et al.* [74]. Spatio-temporal features based on B-splines are extracted in the optical flow field. To model this descriptor, a Bag-of-Words (BoW) technique is employed, whereas, classification of activities is performed using relevant vector machines (RVM) [75].

The classification of a video sequence using local features in a spatio-temporal environment has also been given much focus. Schuldt *et al.* [2] represented local events in a video using space-time features, while an SVM classifier was used to recognize an action. Gorelick *et al.* [76] considered actions as 3D space-time silhouettes of moving humans. They took advantage of the Poisson equation solution to efficiently describe an action by using spectral clustering between sequences of features and applying nearest neighbor classification to characterize an action. Niebles *et al.* [68] addressed the problem of action recognition by creating a codebook of space-time interest points. A hierarchical approach was followed by Jhuang *et al.* [67], where an input video was analyzed into several feature descriptors depending on their complexity. The final classification was performed by a multi-class SVM classifier. Dollár *et al.* [77] proposed spatio-temporal features based on cuboid descriptors. Instead of encoding human motion for action classification, Jainy *et al.* [19] proposed to incorporate information from human-to-objects interactions and combined several datasets to transfer information from one dataset to another.

An action descriptor of histograms of interest points, relying on the work of Schuldt *et al.* [2], was presented by Yan and Luo [78]. Random forests for action representation have also attracted widespread interest for action recognition [79, 80]. Furthermore, the key issue of how many frames are required to recognize an action was addressed by Schindler and Gool [81]. Shabani *et al.* [45] proposed a temporally asymmetric filtering for feature detection and activity recognition. The extracted features were more robust under geometric transformations than the features described by a Gabor filter [82]. Sapienza *et al.* [83] used a bag of local spatio-temporal volume features approach to recognize and localize human actions from weakly labeled video sequences using multiple instance learning.

The problem of identifying multiple persons simultaneously and performing action

recognition was presented by Khamis *et al.* [84]. The authors considered that a person could first be detected by performing background subtraction techniques. Based on histograms of oriented Gaussians, Dalal and Triggs [85] were able to detect humans, whereas classification of actions was made by training an SVM classifier. Wang *et al.* [86] performed human activity recognition by associating the context between interest points based on the density of all features observed. A multi-view activity recognition method was presented by Li and Zickler [46], where descriptors from different views were connected together to construct a new augmented feature that contains the transition between the different views. Multi-view action recognition has also been studied by Rahmani and Mian [87]. A non-linear knowledge transfer model based on deep learning was proposed for mapping action information from multiple camera views into one single view. However, their method is computationally expensive as it requires a two step sequential learning phase prior to the recognition step for analyzing and fusing the information of multi-views.

Tian *et al.* [88] employed spatio-temporal volumes using a deformable part model to train an SVM classifier for recognizing sport activities. Similar in spirit, the work of Jain *et al.* [89] used a 3D space-time volume representation of human actions obtained from super-voxels to understand sport activities. They used an agglomerative approach to merge super-voxels that share common attributes and localize human activities. Kulkarni *et al.* [90] used a dynamic programming approach to recognize sequences of actions in untrimmed video sequences. A per-frame time-series representation of each video and a template representation of each action were proposed, whereas dynamic time warping was used to sequence alignment.

Samanta and Chanda [91] proposed a novel representation of human activities using a combination of spatio-temporal features and a facet model [92], while they used a 3D Haar wavelet transform and higher order time derivatives to describe each interest point. A vocabulary was learned from these features and SVM was used for classification. Jiang *et al.* [93] used a mid-level feature representation of video sequences using optical flow features. These features were clustered using K-means to build a hierarchical template tree representation of each action. A tree search algorithm was used to identify and localize the corresponding activity in test videos. Roshtkhari and Levine [94] also proposed a hierarchical representation of video sequences for recognizing atomic actions by building a codebook of spatio-temporal volumes. A probe video sequence was classified into its underlying activity according to its similarity with each representation in the codebook.

Earlier approaches were based on describing actions by using dense trajectories. The work of Le *et al.* [95] discovered the action label in an unsupervised manner by learning features directly from video data. A high-level representation of video sequences, called "action bank", was presented by Sadanand and Corso [96]. Each video was represented by a set of action descriptors, which were put in correspondence. The final classification was performed by an SVM classifier. Yan and Luo [78] also proposed a novel action descriptor based on spatial temporal interest points (STIP) [97]. To avoid overfitting they proposed

a novel classification technique combining Adaboost and sparse representation algorithms. Wu *et al.* [98] used visual features and Gaussian mixture models (GMM) [25] to efficiently represent the spatio-temporal context distributions between the interest points at several space and time scales. The underlying activity was represented by a set of features extracted by the interest points over the video sequence. A new type of feature called the "hankelet" was presented by Li *et al.* [47]. This type of feature, which was formed by short tracklets, along with a BoW approach, was able to recognize actions under different viewpoints without requiring any camera calibration.

The work of Vrigkas *et al.* [64] focused on recognizing human activities by representing a human action with a set of clustered motion trajectories. A Gaussian mixture model was used to cluster the motion trajectories and the action labeling was performed using a nearest neighbor classification scheme. Yu *et al.* [99] proposed a propagative point matching approach using random projection trees, which can handle unlabeled data in an unsupervised manner. Jain *et al.* [100] used motion compensation techniques to recognize atomic actions. They also proposed a new motion descriptor called "divergence-curl-shear descriptor", which is able to capture the hidden properties of flow patterns in video sequences. Wang *et al.* [15] used dense optical flow trajectories to describe the kinematics of motion patterns in video sequences. However, several intra-class variations caused by missing data, partial occlusion, and the sort duration of actions in time may harm the recognition accuracy. Ni *et al.* [21] discovered the most discriminative groups of similar dense trajectories for analyzing human actions. Each group was assigned a learned weight according to its importance in motion representation.

An unsupervised method for learning human activities from short tracklets was proposed by Gaidon *et al.* [101]. They used a hierarchical clustering algorithm to represent videos with an unordered tree structure and compared all tree-clusters to identity the underlying activity. Raptis *et al.* [63] proposed a mid-level approach extracting spatio-temporal features and constructing clusters of trajectories, which could be considered as candidates of an action. Yu and Yuan [102] extracted bounding box candidates from video sequences, where each candidate may contain human motion. The most significant action paths were estimated by defining an action score. Due to the large spatio-temporal redundancy in videos, many candidates may overlap. Thus, estimation of the maximum set coverage was applied to address this problem. However, the maximum set coverage problem is NP-hard, and thus the estimation requires approximate solutions.

An approach that exploits the temporal information encoded in video sequences was introduced by Li *et al.* [103]. The temporal data were encoded into a trajectory system, which measures the similarity between activities and computes the angle between the associated subspaces. A method that tracks features and produces a number of trajectory snippets was proposed by Matikainen *et al.* [62]. The trajectories were clustered by an SVM classifier. Motion features were extracted from a video sequence by Messing *et al.* [104]. These features were tracked with respect to their velocities and a generative mixture model was employed to learn the velocity history of these trajectories and classify each

video clip. Tran et al. [105] proposed a scale and shape invariant method for localizing complex spatio-temporal events in video sequences. Their method was able to relax the tight constraints of bounding box tracking, while they used a sliding window technique to track spatio-temporal paths maximizing the summation score.

An algorithm that may recognize human actions in 3D space by a multi-camera system was introduced by Holte *et al.* [106]. It was based on the synergy of 3D space and time to construct a 4D descriptor of spatial temporal interest points and a local description of 3D motion features. The BoW technique was used to form a vocabulary of human actions, whereas agglomerative information bottleneck and SVM were used for action classification. Zhou and Wang [107] proposed a new representation of local spatio-temporal cuboids for action recognition. Low level features were encoded and classified via a kernelized SVM classifier, whereas a classification score denoted the confidence that a cuboid belongs to an atomic action. The new feature could act as complementary material to the low level feature. The work of Sanchez-Riera *et al.* [108] recognized human actions using stereo cameras. Based on the technique of BoW, each action was presented by a histogram of visual words, whereas their approach was robust to background clutter.

The problem of temporal segmentation and event recognition was examined by Hoai *et al.* [109]. Action recognition was performed by a supervised learning algorithm. Satkin and Hebert [110] explored the effectiveness of video segmentation by discovering the most significant portions of videos. In the sense of video labeling, the study of Wang *et al.* [111] leveraged the shared structural analysis for activity recognition. The correct annotation was given in each video under a semi supervised scheme. Bag-of-video words have become very popular. Chakraborty *et al.* [112] proposed a novel method applying surround suppression. Human activities were represented by bag-of-video words constructed from spatial temporal interest points by suppressing the background features and building a vocabulary of visual words. Guha and Ward [113] employed a technique of sparse representations for human activity recognition. An overcomplete dictionary was constructed using a set of spatio-temporal descriptors. Classification over three different dictionaries was performed.

Seo and Milanfar [114] proposed a method based on space-time locally adaptive regression kernels and the matrix cosine measure. They extracted features from space-time descriptors and compared them against features of the target video. A vocabulary based approach has been proposed by Kovashka and Grauman [115]. The main idea is to find the neighboring features around the detected interest points, quantize them, and form a vocabulary. Ma *et al.* [116] extracted spatio-temporal segments from video sequences that correspond to whole or part human motion and constructed a tree-structured vocabulary of similar actions. Fernando *et al.* [117] learned to arrange human actions in chronological order in an unsupervised manner by exploiting temporal ordering in video sequences. Relevant information was summarized together through a ranking learning framework.

The main disadvantage of using a global representation such as optical flow is the sensitivity to noise and partial occlusions. Space-time approaches can hardly recognize

actions when more than one person is present in a scene. Nevertheless, space-time features focus mainly on local spatiotemporal information. Moreover, the computation of these features produces sparse and varying numbers of detected interest points, which may lead to low repeatability. However, background subtraction can help overcome this limitation.

Low-level features usually used with a fixed length feature vector (e.g., Bag-of-Words) failed to be associated with high-level events. Trajectory-based methods face the problem of human body detection and tracking, as these are still open issues. Complex activities are more difficult to recognize when space-time feature based approaches are employed. Furthermore, viewpoint invariance is another issue that these approaches have difficulty in handling.

2.3.2 Stochastic Methods

In recent years there has been a tremendous growth in the amount of computer vision research aimed at understanding human activity. There has been an emphasis on activities, where the entity to be recognized may be considered as a stochastically predictable sequence of states. Researchers have conceived and used many stochastic techniques, such as hidden Markov model (HMMs) [25] or hidden conditional random fields (HCRFs) [27], to infer useful results for human activity recognition.

Robertson and Reid [118] modeled human behavior as a stochastic sequence of actions. Each action was described by a feature vector, which combines information about position, velocity, and local descriptors. An HMM was employed to encode human actions, whereas recognition was performed by searching for image features that represent an action. Pioneering this task, Wang and Mori [119] were among the first to propose HCRFs for the problem of activity recognition. A human action was modeled as a configuration of parts of image observations. Motion features were extracted forming a BoW model. Activity recognition and localization via a figure-centric model was presented by Lan *et al.* [49]. Human location was treated as a latent variable, which was extracted from a discriminative latent variable model by simultaneous recognition of an action. A real time algorithm that models human interactions was proposed by Oliver *et al.* [120]. The algorithm was able to detect and track a human movement, forming a feature vector that describes the motion. This vector was given as input to an HMM, which was used for action classification. Song *et al.* [121] considered that human action sequences of various temporal resolutions. At each level of abstraction, they learned a hierarchical model with latent variables to group similar semantic attributes of each layer.

A multi-view person identification was presented by Iosifidis *et al.* [50]. Fuzzy vector quantization and linear discriminant analysis were employed to recognize a human activity. Huang *et al.* [122] presented a boosting algorithm called LatentBoost. The authors trained several models with latent variables to recognize human actions. A stochastic modeling of human activities on a shape manifold was introduced by Yi *et al.* [123]. A human activity was extracted as a sequence of shapes, which is considered as one realization of a random process on a manifold. The piecewise Brownian motion was used to model human activity

on the respective manifold. Wang *et al.* [61] proposed a semi-supervised framework for recognizing human actions combining different visual features. All features were projected onto a common subspace and a boosting technique was employed to recognize human actions from labeled and unlabeled data. Yang *et al.* [20] proposed an unsupervised method for recognizing motion primitives for human action classification from a set of very few examples.

Sun and Nevatia [124] treated video sequences as sets of short clips rather than a whole representation of actions. Each clip corresponded to a latent variable in an HMM model, while a Fisher kernel technique [125] was employed to represent each clip with a fixed length feature vector. Ni *et al.* [126] decomposed the problem of complex activity recognition into two sequential sub-tasks with increasing granularity levels. First, the authors applied human-to-object interaction techniques to identify the area of interest, then used this context-based information to train a conditional random field (CRF) model [26] and identify the underlying action. Lan *et al.* [127] proposed a hierarchical method for predicting future human actions, which may be considered as a reaction to a previous performed action. They introduced a new representation of human kinematic states, called "hierarchical movements", computed at different levels of coarse to fine-grained level granularity. Predicting future events from partially unseen video clips with incomplete action execution has also been studied by Kong *et al.* [128]. A sequence of previously observed features was used as a global representation of actions and a CRF model was employed to capture the evolution of actions across time in each action class.

An approach for group activity classification was introduced by Choi *et al.* [129]. The authors were able to recognize activities such as a group of people talking or standing in a queue. The proposed scheme was based on random forests, which could select samples of spatio-temporal volumes in a video that characterize an action. A probabilistic Markov random field (MRF) [130] framework was used to classify and localize the activities in a scene. Lu *et al.* [131] also employed a hierarchical MRF model to represent segments of human actions by extracting supervoxels from different scales and automatically estimated the foreground motion using saliency features of neighboring supervoxels.

The work of Wang *et al.* [132] focused on tracking dense sample points from video sequences using optical flow based on HCRFs for object recognition. Wang et al. [133] proposed a probabilistic model of two components. The first component modeled the temporal transition between action primitives to handle large variation in an action class, while the second component located the transition boundaries between actions. A hierarchical structure, which is called the sum-product network, was used by Amer and Todorovic [134]. The BoW technique encoded the terminal nodes, the sum nodes corresponded to mixtures of different subsets of terminals, and the product nodes represented mixtures of components.

Zhou and Zhang [135] proposed a robust to background clutter, camera motion and occlusions method for recognizing complex human activities. They used multiple-instance formulation in conjunction with an MRF model and were able to represent human activi-

ties with a bag of Markov chains obtained from STIP and salient region feature selection. Chen *et al.* [136] addressed the problem of identifying and localizing human actions using CRFs. The authors were able to distinguish between intentional actions and unknown motions that may happen in the surroundings by ordering video regions and detecting the actor of each action. Kong and Fu [137] addressed the problem of human interaction classification from subjects that lie close to each other. Such a representation may be erroneous to partial occlusions and feature-to-object mismatching. To overcome this problem the authors proposed a patch-aware model, which learned regions of interacting subjects at different patch levels.

Shu *et al.* [138] recognized complex video events and group activities from aerial shoots captured from unmanned aerial vehicles (UAVs). A preprocessing step prior to the recognition process was adopted to address several limitations of frame capturing such as low resolution, camera motion, and occlusions. Complex events were decomposed into simpler actions and modeled using a spatiotemporal CRF graph. A video segmentation approach for video activities and a decomposition into smaller clips task that contained sub-actions was presented by Wu *et al.* [139]. The authors modeled the relation of consecutive actions by building a graphical model for unsupervised learning of the activity label from depth sensor data.

Often, human actions are highly correlated to the actor, who performs a specific action. Understanding both the actor and the action may be vital for real life applications such as robot navigation or patient monitoring. Most of the existing works do not take into account the fact that a specific action may be performed in different manner by a different actors. Thus, a simultaneous inference of actors and actions is required. Xu *et al.* [140] addressed these limitations and proposed a general probabilistic framework for joint actor-action understanding while they presented a new dataset for actor-action recognition.

There is an increasing interest in exploring human-object interaction for recognition. Moreover, recognizing human actions from still images by taking advantage of contextual information such as surrounding objects is a very active topic [141]. These methods assume that not only the human body itself, but the objects surrounding it, may provide evidence of the underlying activity. For example, a soccer player interacts with a ball when playing soccer. Motivated by this fact, Gupta and Davis [11] proposed a Bayesian approach that encodes object detection and localization for understanding human actions. Extending the previous method, Gupta *et al.* [142] introduced spatial and functional constraints on static shape and appearance features and they were also able to identify human-to-object interactions without incorporating any motion information. Ikizler-Cinbis and Sclaroff [143] extracted dense features and performed tracking over consecutive frames for describing both motion and shape information. Instead of explicitly using separate object detectors, they divided the frames into regions and treated each region as an object candidate.

Most of the existing probabilistic methods for human activity recognition may perform well and apply exact and/or approximate learning and inference. However, they are

usually more complicated than non-parametric methods, since they use dynamic programming or computationally expensive HMMs for estimating a varying number of parameters. Due to their Markovian nature, they must enumerate all possible observation sequences while capturing the dependencies between each state and its corresponding observation only. HMMs treat features as conditionally independent, but this assumption may not hold for the majority of applications. Often, the observation sequence may be ignored due to normalization leading to the label bias problem [26]. Thus, HMMs are not suitable for recognizing more complex events, but rather an event is decomposed into simpler activities, which are easier to recognize.

CRFs on the other hand, overcome the label bias problem. Most of the aforementioned methods do not require large training datasets, since they are able to model the hidden dynamics of the training data and incorporate prior knowledge over the representation of data. Although CRFs outperform HMMS in many applications, including bioinformatics, activity, or speech recognition, the construction of more complex models for human activity recognition may have a good generalization ability, but is rather impractical for real time applications due to the large number of parameter estimations and the approximate inference.

2.3.3 Rule-Based Methods

Rule based approaches determine ongoing events by modeling an activity using rules or sets of attributes that describe an event. Each activity is considered as a set of primitive rules/attibutes, which enables the construction of a descriptive model for human activity recognition.

Action recognition of complex scenes with multiple subjects was proposed by Morariu and Davis [51]. Each subject must follow a set of certain rules while performing an action. The recognition process was performed over basketball game videos, where the players were first detected and tracked, generating a set of trajectories that are used to create a set of spatio-temporal events. Based on first order logic and probabilistic approaches such as Markov networks, the authors were able to infer which event has occurred. Liu *et al.* [144] addressed the problem of recognizing actions by a set of descriptive and discriminative attributes. Each attribute was associated with the characteristics describing the spatio-temporal nature of the activities. These attributes were treated as latent variables, which capture the degree of importance of each attribute for each action in a latent SVM approach.

A combination of activity recognition and localization was presented by Chen and Grauman [52]. The whole approach was based on the construction of a space-time graph using a high-level descriptor, where the algorithm seeks to find the optimal subgraph that maximizes the activity classification score (i.e., find the maximum weight subgraph, which in the general case is an NP-complete problem). Kuehne *et al.* [145] proposed a structured temporal approach for daily living human activity recognition. The author used HMMs to model human actions as action units and then used grammatical rules to form a sequence

of complex actions by combining different action units. When temporal grammars are used for action classification, the main problem consists in treating long video sequences due to the complexity of the models. One way to cope with this limitation is to segment video sequences into smaller clips that contain sub-actions, using a hierarchical approach [146]. The generation of short description from video sequences [147] based on convolutional neural networks (CNN) [148] was also used for activity recognition [149].

Intermediate semantic features representation for recognizing unseen actions during training were proposed [150]. These intermediate features were learned during training, while parameter sharing between classes was enabled by capturing the correlations between frequently occurring low-level features [151]. Learning how to recognize new classes that were not seen during training, by associating intermediate features and class labels, is a necessary aspect for transferring knowledge between training and test samples. This problem is generally known as zero-shot learning [152]. Thus, instead of learning one classifier per attribute, a two step classification method has been proposed by Lampert *et al.* [153]. Specific attributes are predicted from already learned classifiers and are mapped into a class-level score.

Action classification from still images by learning semantic attributes was proposed by Yao *et al.* [154]. Attributes describe specific properties of human actions, while parts of actions, which were obtained from objects and human poses, were used as bases for learning complex activities. The problem of attribute-action association was reported by Zhang *et al.* [155]. The authors proposed a multi-task learning approach [156] for simultaneously coping with low-level features and action-attribute relationships, and introduced attribute regularization as a penalty term for handling irrelevant predictions. A robust to noise representation of attribute-based human action classification was proposed by Zhang *et al.* [157]. Sigmoid and Gaussian envelopes were incorporated into the loss function of an SVM classifier, where the outliers are eliminated during the optimization process. A GMM was used for modeling human actions and a transfer ranking technique was employed for recognizing unseen classes. Ramanathan *et al.* [158] were able to transfer semantic knowledge between classes to learn human actions from still images. The interaction between different classes was performed using linguistic rules. However, for high-level activities the use of language priors is often not adequate, thus simpler and more explicit rules should be constructed.

Complex human activities cannot be recognized directly from rule-based approaches. Thus, decomposition into simpler atomic actions is applied and then combination of individual actions is employed for the recognition of complex or simultaneously occurring activities. This limitation leads to constant feedback by the user of rule/attribute annotations of the training examples, which is time consuming and sensitive to errors due to subjective point of view of the user defined annotations. To overcome this drawback, several approaches employing transfer learning [153, 159], multi-task learning [156, 160], and semantic/discriminative attribute learning [161, 162] were proposed to automatically generate and handle the most informative attributes for human activity classification.

2.3.4 Shape-Based Methods

Modeling of human pose and appearance has received a great response from researchers during the last decades. Parts of the human body are described in 2D space as rectangular patches and as volumetric shapes in 3D space. It is well known that activity recognition algorithms based on the human silhouette play an important role in recognizing human actions. As a human silhouette consists of limbs jointly connected to each other, it is important to obtain exact human body parts from videos. This problem is considered as part of the action recognition process. Many algorithms convey a wealth of information about solving this problem.

A major focus in action recognition from still images or videos has been made in the context of scene appearance [163, 164, 165]. More specifically, Thurau and Hlavac, [163] represented actions by histograms of pose primitives and n-gram expressions were used for action classification. Also, Yang *et al.* [164] combined actions and human poses together, treating poses as latent variables, to infer the action label in still images. Maji *et al.* [165] introduced a representation of human poses, called the "poselet activation vector", which is defined by the 3D orientation of the head and torso and provided a robust representation of human pose and appearance. Moreover, action categorization based on modeling the motion of parts of the human body was presented by Tran *et al.* [53], where a sparse representation was used to model and recognize complex actions. In the sense of template matching techniques, Rodriguez *et al.* [3] introduced the maximum average correlation height (MACH) filter, which was a method for capturing intra-class variabilities by synthesizing a single action MACH filter for a given action class. Sedai *et al.* [166] proposed a combination of shape and appearance descriptors to represent local features for human pose estimation. The different types of descriptors were fused at the decision level using a discriminative learning model. Nevertheless, identifying which body parts are most significant for recognizing complex human activities still remains a challenging task [167].

İkizler and Duygulu [168] modeled the human body as a sequence of oriented rectangular patches. The authors described a variation of BoW method called bag-of-rectangles. Spatially oriented histograms were formed to describe a human action, while the classification of an action was performed using four different methods such as frame voting, global histogramming, SVM classification, and dynamic time warping (DTW) [169]. The study of Yao and Fei-Fei [170] modeled human poses for human-object interactions by introducing a mutual context model. The types of human poses, as well as the spatial relationship between the different human parts, were modeled. Self organizing maps (SOM) [171] were introduced by Iosifidis *et al.* [172] for learning human body posture, in conjunction with fuzzy distances, to achieve time invariant action representation. The proposed algorithm was based on multi-layer perceptrons, where each layer was fed by an associated camera, for view-invariant action classification. Human interactions were addressed by Andriluka and Sigal [173]. First, 2D human poses were estimated from pictorial structures from groups of humans and then each estimated structure was fitted

into 3D space. To this end, several 2D human pose benchmarks have been proposed for the evaluation of articulated human pose estimation methods [174].

Action recognition using depth cameras was introduced by Wang *et al.* [175], where a new feature type called "local occupancy pattern" was also proposed. This feature was invariant to translation and was able to capture the relation between human body parts. The authors also proposed a new model for human actions called "actionlet ensemble model", which captured the intra-class variations and was robust to errors incurred by depth cameras. 3D human poses have been taken into consideration in recent years and several algorithms for human activity recognition have been developed. A recent review on 3D pose estimation and activity recognition was proposed by Holte *et al.* [176]. The authors categorized 3D pose estimation approaches aimed at presenting multi-view human activity recognition methods. The work of Shotton *et al.* [177] modeled 3D human poses and performed human activity recognition from depth images by mapping the pose estimation problem into a simpler pixel-wise classification problem. Graphical models have been widely used in modeling 3D human poses. The problem of articulated 3D human pose estimation was studied by Fergie and Galata [178], where the limitation of the mapping from the image feature space to the pose space was addressed using mixtures of Gaussian processes, particle filtering, and annealing [179]. A combination of discriminative and generative models improved the estimation of human pose.

Multi-view pose estimation was examined by Amin *et al.* [180]. The 2D poses for different sources were projected onto 3D space using a mixture of multi-view pictorial structures models. Belagiannis *et al.* [181] have also addressed the problem of multi-view pose estimation. They constructed 3D body part hypotheses by triangulation of 2D pose detections. To solve the problem of body part correspondence between different views, the authors proposed a 3D pictorial structure representation based on a CRF model. However, building successful models for human pose estimation is not straightforward [182]. Combining both pose specific appearance and the joint appearance of body parts helps to construct a more powerful representation of the human body. Deep learning has gained much attention for multi-source human pose estimation [183] where the tasks of detection and estimation of human pose were jointly learned. Toshev and Szegedy [184] have also used deep learning for human pose estimation. Their approach relies on using deep neural networks (DNN) [185] for representing cascade body joint regressors in a holistic manner.

Despite the vast development of pose estimation algorithms, the problem still remains challenging for real time applications. Jung *et al.* [186] presented a method for fast estimation of human pose with 1,000 frames per second. To achieve such a high computational speed the authors used random walk sub-sampling methods. Human body parts were handled as directional tree-structured representations and a regression tree was trained for each joint in the human skeleton. However, this method depends on the initialization of the random walk process.

Sigal *et al.* [54] addressed the multi-view human tracking problem where the modeling

of 3D human pose consisted of a collection of human body parts. The motion estimation was performed by non-parametric belief propagation [25]. On the other hand, the work of Livne *et al.* [187] explored the problem of inferring human attributes, such as gender, weight, and mood, by the scope of 3D pose tracking. Representing activities using trajectories of human poses is computationally expensive due to many degrees of freedom. To this end, efficient dimensionality reduction methods should be applied. Moutzouris *et al.* [188] proposed a novel method for reducing dimensionality of human poses called "hierarchical temporal Laplacian eigenmaps" (HTLE). Moreover, the authors were able to estimate unseen poses using a hierarchical manifold search method.

Du *et al.* [189] divided the human skeleton into five segments and used each of these parts to train a hierarchical neural network. The output of each layer, which corresponds to neighboring parts, is fused and fed as input to the next layer. However, this approach suffers from the problem of data association as parts of the human skeleton may vanish through the sequential layer propagation and back projection. Nie *et al.* [190] also divided human pose into smaller mid-level spatio-temporal parts. Human actions were represented using a hierarchical AND/OR graph and dynamic programming was used to infer the class label. One disadvantage of this method is that it cannot deal with self-occlusions (i.e., overlapping parts of human skeleton).

A shared representation of human poses and visual information has also been explored [7, 191, 192]. However, the effectiveness of such methods is limited by tracking inaccuracies in human poses and complex backgrounds. To this end, several kinematic and part-occlusion constraints for decomposing human poses into separate limbs have been explored to localize the human body [193]. Xu *et al.* [194] proposed a mid-level representation of human actions by computing local motion volumes in skeletal points extracted from video sequences, and constructed a codebook of poses for identifying the action. Eweiwi *et al.* [195] reduced the required amount of pose data using a fixed length vector of more informative motion features (e.g., location and velocity) for each skeletal point. A partial least squares approach was used for learning the representation of action features, which is then fed into an SVM classifier.

Kviatkovsky *et al.* [196] mixed shape and motion features for online action classification. The recognition processes could be applied in real time using the incremental covariance update and the on-demand nearest neighbor classification schemes. Rahmani *et al.* [197] trained a random decision forest (RDF) [198] and applied a joint representation of depth information and 3D skeletal positions for identifying human actions in real time. A novel part-based skeletal representation for action recognition was introduced by Vemulapalli *et al.* [199]. The geometry between different body parts was taken into account and a 3D representation of human skeleton was proposed. Human actions are treated as curves in the Lie group [200] and the classification was performed using SVM and temporal modeling approaches. Following a similar approach, Anirudh *et al.* [201] represented skeletal joints as points on the product space. Shape features were represented as high dimensional non-linear trajectories on a manifold to learn the latent variable space

of actions. Fouhey *et al.* [202] exploited the interaction between human actions and scene geometry to recognize human activities from still images using 3D skeletal representation and adopting geometric representation constraints of the scenes.

The problem of appearance-to-pose mapping for human activity understanding was studied by Urtasun and Darrell [203]. Gaussian processes were used as an online probabilistic regressor for this task using sparse representation of data for reducing computational complexity. Theodorakopoulos *et al.* [204] have also employed sparse representation of skeletal data in the dissimilarity space for human activity recognition. In particular, human actions are represented by vectors of dissimilarities and a set of prototype actions is built. The recognition is performed into the dissimilarity space using sparse representation-based classification. A publicly available dataset (UPCV Action dataset) consisting of skeletal data of human actions was also proposed.

A common problem in estimating human pose is the high-dimensional space (i.e., each limb may have a large number of degrees of freedom that need to be estimated simultaneously). Action recognition relies heavily on the obtained pose estimations. The articulated human body is usually represented as a tree-like structure, thus locating the global position and tracking each limb separately is intrinsically difficult, since it requires exploration of a large state space of all possible translations and rotations of the human body parts in 3D space. Many approaches, which employ background subtraction [205] or assume fixed limb lengths and uniformly distributed rotations of body parts [206], have been proposed to reduce the complexity of the 3D space.

Moreover, the association of human pose orientation with the poses extracted from different camera views is also a difficult problem due to similar body parts of different humans in each view. Mixing body parts of different views may lead to ambiguities because of the multiple candidates of each camera view and false positive detections. The estimation of human pose is also very sensitive to several factors such as illumination changes, variations in view-point, occlusions, background clutter, and human clothing. Low cost devices such as Microsoft Kinect or other RGB-D sensors, which provide 3D depth data of a scene, can efficiently leverage these limitations and produce a relatively good estimation of human pose, since they are robust to illumination changes and texture variations [207].

2.4 Multimodal Methods

Recently, much attention has been focused on multimodal activity recognition methods. An event can be described by different types of features that provide more and useful information. In this context, several multimodal methods are based on feature fusion, which can be expressed by two different strategies: early fusion and late fusion. The easiest way to gain the benefits of multiple features is to directly concatenate features in a larger feature vector and then learn the underlying action [208]. This feature fusion technique may improve recognition performance, but the new feature vector is of much

larger dimension.

Multimodal cues are usually correlated in time, thus a temporal association of the underlying event and the different modalities is an important issue for understanding the data. In that context, audio-visual analysis is used in many applications, not only for audio-visual synchronization [209], but also for tracking [210] and activity recognition [55]. Multimodal methods are classified into three categories: (i) *affective methods*, (ii) *behavioral methods*, and (iii) *methods based on social networking*. Multimodal methods describe atomic actions or interactions that may correspond to affective states of a person with whom he/she communicates, and depend on emotions and/or body movements.

2.4.1 Affective Methods

The core of emotional intelligence is understanding the mapping between a person's affective states and the corresponding activities, which are strongly related to the emotional state and communication of a person with other people [211]. Affective computing studies model the ability of a person to express, recognize, and control his/her affective states in terms of hand gestures, facial expressions, physiological changes, speech, and activity recognition [40]. This research area is generally considered to be a combination of computer vision, pattern recognition, artificial intelligence, psychology, and cognitive science.

A key issue in affective computing is accurately annotated data. Ratings are one of the most popular affect annotation tools. However, this is challenging to obtain for real world situations, since affective events are expressed in a different manner by different persons, or occur simultaneously with other activities and feelings. Preprocessing affective annotations may be detrimental for generating accurate and ambiguous affective models due to biased representations of affect annotation. To this end, a study on how to produce highly informative affective labels has been proposed by Healey [212]. Soleymani *et al.* [213] investigated the properties of developing a user-independent emotion recognition system that is able to detect the most informative affective tags from electroencephalogram (EEG) signals, pupillary reflex, and bodily responses that correspond to video stimulus. Nicolaou *et al.* [214] proposed a novel method based on probabilistic canonical correlation analysis (PCCA) [215] and DTW for fusing multimodal emotional annotations and performing temporal aligning of sequences.

Liu *et al.* [56] associated multimodal features (i.e., textual and visual) for classifying affective states in still images. The authors argued that visual information is not adequate for understanding human emotions, and thus additional information that describes the image is needed. Dempster-Shafer theory [216] was employed for fusing the different modalities, while SVM was used for classification. Hussain *et al.* [217] proposed a framework for fusing multimodal psychological features such as heart and facial muscle activity, skin response, and respiration, for detecting and recognizing affective states. Al-Zoubi *et al.* [218] explored the effect of the affective feature variations over time on the classification of affective states.

Siddiquie *et al.* [219] analyzed four different affective dimensions such as activation,

expectancy, power, and valence [220]. To this end, they proposed joint hidden conditional random Fields (JHCRF) as a new classification scheme to take advantage of the multimodal data. Furthermore, their method uses late fusion to combine audio and visual information together. This may lead to significant loss of the inter-modality dependence, while it suffers from propagating the classification error across different levels of classifiers. Although their method could efficiently recognize the affective state of a person, the computational burden was high as JHCRFs require twice as many hidden variables as the traditional HCRFs when features represent two different modalities.

Nicolaou *et al.* [221] proposed a regression model based on SVMs for regression (SVR) [222] for continuous prediction of multimodal emotional states, using facial expression, shoulder gesture, and audio cues in terms of arousal and valence. Castellano *et al.* [59] explored the dynamics of body movements to identify affective behaviors using time series of multimodal data. Martinez *et al.* [23] presented a detailed review of learning methods for classification of affective and cognitive states of computer game players. They analyzed the properties of directly using affect annotations in classification models, and proposed a method for transforming such annotations to build more accurate models.

Multimodal affect recognition methods in the context of neural networks and deep learning have generated considerable recent research interest [223]. In a more recent study, Martinez *et al.* [224] could efficiently extract and select the most informative multimodal features using deep learning to model emotional expressions and recognize the affective states of a person. They incorporated psychological signals into emotional states such as relaxation, anxiety, excitement, and fun, and demonstrated that deep learning was able to extract more informative features than feature extraction on psychological signals.

Although the understanding of human activities may benefit from affective state recognition, the classification process is extremely challenging due to the semantic gap between the low-level features extracted from video frames and high-level concepts such as emotions that need to be identified. Thus, building strong models that can cope with multimodal data, such as gestures, facial expressions or psychological data, depends on the ability of the model to discover relations between different modalities and generate informative representation on affect annotations. Generating such information is not an easy task. Users cannot always express their emotion with words, and producing satisfactory and reliable ground truth that corresponds to a given training instance is quite difficult as it can lead to ambiguous and subjective labels. This problem becomes more prominent as human emotions are continuous acts in time and variations in human actions may be confusing or lead to subjective annotations. Therefore, automatic affective recognition systems should reduce the effort for selecting the proper affective label to better assess human emotions.

2.4.2 Behavioral Methods

Recognizing human behaviors from video sequences is a challenging task for the computer vision community [225]. A behavior recognition system may provide information about

the personality and psychological state of a person and its applications vary from video surveillance to human-computer interaction. Behavioral approaches aim at recognizing behavioral attributes, non-verbal multimodal cues such as gestures, facial expressions, and auditory cues. Factors that can affect human behavior may be decomposed into several components including emotions, moods, actions, and interactions with other people. Hence, the recognition of complex actions may be crucial for understanding human behavior. One important aspect of human behavior recognition is the choice of proper features, which can be used to recognize behavior in applications such as gaming or physiology. A key challenge in recognizing human behaviors is to define specific emotional attributes for multimodal dyadic interactions [226]. Such attributes may be descriptions of emotional states or cognitive states such as activation, valence, or engagement.

Audio-visual representation of human actions has gained an important role in human behavior recognition methods. Sargin *et al.* [227] suggested a method for speaker identification integrating a hybrid scheme of early and late fusion of audio-visual features and used CCA [228] to synchronize the multimodal features. However, their method can cope with video sequences of frontal view only. Metallinou *et al.* [229] proposed a probabilistic approach based on GMMs for recognizing human emotions in dyadic interactions. The authors took advantage of facial expressions as they can be expressed by the facial action coding system (FACS) [230], which describes all possible facial expressions as a combination of action units (AU), and combines them with audio information, extracted from each participant, to identify their emotional state. Similarly, Chen *et al.* [231] proposed a real-time emotion recognition system that modeled 3D facial expressions using random forests. The proposed method was robust to subjects' poses and changes in the environment.

Wu *et al.* [232] proposed a human activity recognition system by taking advantage of the auditory information of the video sequences of the HOHA dataset [233] and used late fusion techniques for combining audio and visual cues. The main disadvantage of this method is that it used different classifiers to separately learn the audio and visual context. Also, the audio information of the HOHA dataset contains dynamic backgrounds and the audio signal is highly diverse (i.e., audio shifts roughly from one event to another), which generates the need for developing audio feature selection techniques. Similar in spirit is the work of Wu *et al.* [55], who used the generalized multiple kernel learning algorithm for estimating the most informative audio features. They applied fuzzy integral techniques to combine the outputs of two different SVM classifiers, increasing the computational burden of the method.

Song *et al.* [57] proposed a novel method for human behavior recognition based on multi-view hidden conditional random fields (MV-HCRF) [234] and estimated the interaction of the different modalities by using kernel canonical correlation analysis (KCCA) [228]. However, their method cannot deal with data that contain complex backgrounds, and due to the down-sampling of the original data the audio-visual synchronization may be lost. Also, their method used different sets of hidden states for audio and visual

information. This property considers that the audio and visual features were a priori synchronized, while it increases the complexity of the model. Metallinou *et al.* [235] employed several hierarchical classification models from neural networks to HMMs and their combinations to recognize audio-visual emotional levels of valence and arousal rather than emotional labels such as anger or kindness.

Vrigkas *et al.* [5] employed a fully connected CRF model to identify human behaviors such as friendly, aggressive and neutral. To evaluate their method they introduced a novel behavior dataset, called the *Parliament* dataset, which consists of political speeches in the Greek parliament. Bousmalis *et al.* [236] proposed a method based on hierarchical Dirichlet processes to automatically estimate the optimal number of hidden states in an HCRF model for identifying human behaviors. The proposed model, also known as infinite hidden conditional random field model (iHCRF), was employed to recognize emotional states such as pain and agreement and disagreement from non-verbal multimodal cues.

Baxter *et al.* [237] proposed a human classification model that does not learn the temporal structure of human actions but rather decomposes human actions and uses them as features for learning complex human activities. The intuition behind this approach is a psycholinguistics phenomenon, where randomizing letters in the middle of words has almost no effect on understanding the underlying word if and only if the first and the last letters of this word remain unchanged [238]. The problem of behavioral mimicry in social interactions was studied by Bilakhia *et al.* [239]. It can be seen as an interpretation of human speech, facial expressions, gestures, and movements. Metallinou *et al.* [240] applied mixture models to capture the mapping between audio and visual cues to understand the emotional states of dyadic interactions.

Selecting the proper features for human behavior recognition has always been a trial-and-error approach for many researchers in this area of study. In general, effective feature extraction is highly application dependent. Several feature descriptors such as HOG3D [241] or STIP [97] are not able to sufficiently characterize human behaviors. The combination of visual features with other more informative features, which reflect human emotions and psychology, is necessary for this task. Nonetheless, the description of human activities with high-level contents usually leads to recognition methods with high computational complexity. Another obstacle that researchers must overcome is the lack of adequate benchmark datasets to test and validate the reliability, effectiveness, and efficiency of a human behavior recognition system.

2.4.3 Methods Based on Social Networking

Social interactions are an important part of daily life. A fundamental component of human behavior is the ability to interact with other people via their actions. Social interaction can be considered as a special type of activity where someone adapts his/her behavior according to the group of people surrounding him/her. Most of the social networking systems that affect people's behavior, such as Facebook, Twitter, or YouTube. measure social interactions and infer how such sites may be involved in issues of identity,

privacy, social capital, youth culture, and education. Moreover, the field of psychology has attracted great interest in studying social interactions, as scientists may infer useful information about human behavior. A recent survey on human behavior recognition provides a complete summarization of up-to-date techniques for automatic human behavior analysis for single person, multi-person, and object-person interactions [225].

Fathi *et al.* [242] modeled social interactions by estimating the location and orientation of the faces of persons taking part in a social events, computing a line of sight for each face. This information was used to infer the location where an individual may be found. The type of interaction was recognized by assigning social roles to each person. The authors were able to recognize three types of social interactions: dialogue, discussion, and monologue. To capture these social interactions, eight subjects wearing head-mounted cameras participated in groups of interacting persons analyzing their activities from the first-person point of view. In the sense of first-person scene understanding, Park and Shi [243] were able to predict joint social interactions by modeling geometric relationships between groups of interacting persons. Although the proposed method could cope with missing information and variations in scene context, scale, and orientation of human poses, it is sensitive to localization of interacting members, which leads to erroneous predictions of the true class.

Human behavior on sport datasets was investigated by Lan *et al.* [24]. The authors modeled the behavior of humans in a scene using social roles in conjunction with modeling low-level actions and high-level events. Burgos-Artizzu *et al.* [244] discussed the social behavior of mice. Each video sequence was segmented into periods of activities by constructing a temporal context that combines spatio-temporal features. Kong *et al.* [60] proposed a new high-level descriptor called "interactive phrases" to recognize human interactions. This descriptor was a binary motion relationship descriptor for recognizing complex human interactions. Interactive phrases were treated as latent variables, while the recognition was performed using a CRF model.

Cui *et al.* [245] recognized abnormal behaviors in human group activities. The authors represented human activities by modeling the relationships between the current behavior of a person and his/her actions. An attribute-based social activity recognition method was introduced by Fu *et al.* [246]. The authors were interested in classifying social activities of daily life such as birthdays or weddings. A new social activity dataset has also been proposed. By treating attributes as latent variables, the authors were able to annotate and classify video sequences of social activities. Yan *et al.* [16] leveraged the problem of human tracking for modeling the repulsion, attraction, and non-interaction effects in social interactions. The tracking problem was decomposed into smaller tasks by tracking all possible configurations of interactions effects, while the number of trackers was dynamically estimated. Tran *et al.* [22] modeled crowded scenes as a graph of interacting persons. Each node represents one person and each edge on the graph is associated with a weight according to the level of the interaction between the participants. The interacting groups were found by graph clustering, where each maximal clique corresponds to an

interacting group.

The work of Lu *et al.* [247] focused on automatically tracking and recognizing players' positions (i.e., attacker, defender) in sports videos. The main problem of this work was the low resolution of the players to be tracked (a player was roughly 15 pixels tall). Lan *et al.* [248] recognized group activities, which were considered as latent variables, encoding the contextual information in a video sequence. Two types of contextual information were explored: group-to-person interactions and person-to-person interactions. To model person-to-person interactions, one approach is to model the associated structure. The second approach is based on spatio-temporal features, which encode the information about an action and the behavior of people in the neighborhood. Finally, the third approach is a combination of the above two.

Much focus has also been given to recognizing human activities from real life videos such as movies or TV shows by exploiting scene contexts to localize activities and understand human interactions [6, 249, 250, 251]. The recognition accuracy of such complex videos can also be improved by relating textual descriptions and visual context to a unified framework [252]. An alternative approach is a system that takes a video clip as its input and generates short textual descriptions, which may correspond to an activity label, which was unseen during training [253]. However, natural video sequences may contain irrelevant scenes or scenes with multiple actions. As a result, Bandla and Grauman [254] proposed a method for recognizing human activities from unsegmented videos using a voting-based classification scheme to find the most frequently used action label.

Marín-Jiménez *et al.* [58] used a bag of visual-audio words scheme along with late fusion for recognizing human interactions in TV shows. Even though their method performs well in recognizing human interaction, the lack of an intrinsic audio-visual relationship estimation limits the recognition problem. Bousmalis et al. [255] considered a system based on HCRFs for spontaneous agreement and disagreement recognition using audio and visual features. Although both methods yielded promising results, they did not consider any kind of explicit correlation and/or association between the different modalities. Hoai and Zisserman [251] proposed a learning based method based on the context and the properties of a scene for detecting upper body positions and understanding the interaction of the participants in TV shows. An audio-visual analysis for recognizing dyadic interactions was presented by Yang *et al.* [256]. The author combined a GMM with a Fisher kernel to model multimodal dyadic interactions and predict the body language of each subject according to the behavioral state of his/her interlocutor. Escalera *et al.* [257] represented the concept of social interactions as an oriented graph using an influence model to identify human interactions. Audio and visual detection and segmentation were performed to extract the exact segments of interest in a video sequence, and then the influence model was employed. Each link measured the influence of a person over another.

Many works on human activity recognition based on deep learning techniques have been proposed in the literature. In fact, deep learning methods have had a large impact

on a plethora of research areas including image/video understanding, speech recognition, and biomedical image analysis. Kim *et al.* [258] used deep belief networks (DBN) [259] in both a supervised and unsupervised manner to learn the most informative audio-visual features and classify human emotions in dyadic interactions. Their system was able to preserve non-linear relationships between multimodal features and showed that unsupervised learning can be used efficiently for feature selection. Shao *et al.* [260] mixed appearance and motion features for recognizing group activities in crowded scenes collected from the web. For the combination of the different modalities the authors applied multi-task deep learning. By these means, they were able to capture the intra-class correlations between the learned attributes while they proposed a novel dataset of crowed scene understanding, called WWW crowd dataset.

Deep learning has also been used by Gan *et al.* [18] for detecting and recognizing complex events in video sequences. The proposed approach followed a sequential framework. First, saliency maps were used for detecting and localizing events and then deep learning was applied to the pre-trained features for identifying the most important frames that correspond to the underlying event. Although much of the existing work on event understanding relies on video representation, significant work has been done on recognizing complex events from static images. Xiong *et al.* [261] utilized CNNs to hierarchically combine information from different visual channels. The new representation of fused features was used to recognize complex social events. To assess their method, the authors introduced a large dataset with more than $60,000$ static images obtained from the web, called web image dataset for event recognition (WIDER).

Karpathy *et al.* [262] performed an experimental evaluation of CNNs to classify events from large-scale video datasets, using one million videos with 487 categories (Sports-1M dataset) obtained from YouTube videos. Chen *et al.* [263] exploited different types of features such as static and motion features for recognizing unlabeled events from heterogenous web data (e.g., YouTube, Google/Bing image search engines). A separate classifier for each source is learned and a multi-domain adaptation approach was followed to infer the labels for each data source. Tang *et al.* [264] studied the problem of heterogenous feature combination for recognizing complex events. They considered the problem as two different tasks. At first, they estimated which were the most informative features for recognizing social events, and then combined the different features using an AND/OR graph structure.

Modeling crowded scenes has been a difficult task due to partial occlusions, interacting motion patterns, and sparsely distributed cameras in outdoor environments [265]. Most of the existing approaches for modeling group activities and social interactions between different persons usually exploit contextual information from the scenes. However, such information is not sufficient to fully understand the underlying activity as it does not capture the variations in human poses when interacting with other persons. When attempting to recognize social interactions with a fixed number of participants the problem may become more or less trivial. When the number of interacting people dynamically changes over

time, the complexity of the problem increases and becomes more challenging. Moreover, social interactions are usually decomposed into smaller subsets that contain individual person activities or interaction between individuals. The individual motion patterns are analyzed separately and are then combined to estimate the event. A person adapts his/her behavior according to the person with whom s/he interacts. Thus, such an approach is limited by the fact that only specific interaction patterns can be successfully modeled and is sensitive in modeling complex social events.

2.4.4 Multimodal Feature Fusion

Consider the scenario where several people have a specific activity/behavior and some of them may emit sounds. In the simple case, a human activity recognition system may recognize the underlying activity by taking into account only the visual information. However, the recognition accuracy may be enhanced from audio-visual analysis, as different people may exhibit different activities with similar body movements, but with different sound intensity values. The audio information may help to understand who is the person of interest in a test video sequence and distinguish between different behavioral states.

A great difficulty in multimodal feature analysis is the dimensionality of the data from different modalities. For example, video features are much more complex with higher dimensions than audio, and thus techniques for dimensionality reduction are useful. In the literature, there are two main fusion strategies that can be used to tackle this problem [266, 267].

Early fusion, or fusion at the feature level, combines features of different modalities, usually by reducing the dimensionality in each modality and creating a new feature vector that represents an individual. Canonical correlation analysis (CCA) [228] was widely studied in the literature as an effective way for fusing data at the feature level [268, 269, 270]. The advantage of early fusion is that it yields good recognition results when the different modalities are highly correlated, since only one learning phase is required. On the other hand, the difficulty of combining the different modalities may lead to the domination of one modality over the others. A novel method for fusing verbal (i.e., textual information) and non-verbal (i.e., visual signals) cues was proposed by Evangelopoulos *et al.* [271]. Each modality is separately analyzed and saliency scores are used for linear and non-linear fusing schemes.

The second category of methods, which is known as *late fusion* or fusion at the decision level, combines several probabilistic models to learn the parameters of each modality separately. Then all scores are combined together in a supervised framework yielding a final decision score [272, 273]. The individual strength of each modality may lead to better recognition results. However, this strategy is time-consuming and requires more complex supervised learning schemes, which may cause a potential loss of inter-modality correlation. A comparison of early versus late fusion methods for video analysis was reported by Snoek *et al.* [274].

Recently, a third approach for fusing multimodal data has come to the foreground

[262]. This approach, called *slow fusion*, is a combination of the previous approaches and can be seen as a hierarchical fusion technique that slowly fuses data by successively passing information through early and late fusion levels. Although this approach seems to have the advantages of both early and late fusion techniques, it also has a large computational burden due to the different levels of information processing.

2.5 Discussion

Human activity understanding has become one of the most active research topics in computer vision. The type and amount of data that each approach uses depends on the ability of the underlying algorithm to deal with heterogeneous and/or large scale data. The development of a fully automated human activity recognition system is a non-trivial task due to cluttered backgrounds, complex camera motion, large intra-class variations, and data acquisition issues. Tables 2.2 and 2.3 provide a comprehensive comparison of unimodal and multimodal methods, respectively, and list the benefits and limitations of each family of methods.

The first step in developing a human activity recognition system is to acquire an adequate human activity database. This database may be used for training and testing purposes. A complete survey, which covers important aspects of human activity recognition datasets, was introduced by Chaquet *et al.* [275]. An appropriate human activity dataset is required for the development of a human activity recognition system. This dataset should be sufficiently rich in a variety of human actions. Moreover, the creation of such a dataset should correspond to real world scenarios. The quality of the input media that forms the dataset is one of the most important things one should take into account. These input media can be static images or video sequences, colored or gray-scaled. An ideal human activity dataset should address the following issues: (i) the input media should include either still images and/or video sequences, (ii) the amount of data should be sufficient, (iii) input media quality (resolution, grayscale or color), (iv) large number of subjects performing an action, (v) large number of action classes, (vi) changes in illuminations, (vii) large intra-class variations (i.e., variations in subjects' poses), (viii) photo shooting under partial occlusion of human structure, and (ix) complex backgrounds.

Although there exists a plethora of benchmark activity recognition datasets in the literature, we have focused on the most widely used ones with respect to the database size, resolution, and usability. Table 2.4 summarizes human activity recognition datasets, categorizing them into seven different categories. All datasets are grouped by their associated category and by chronological order for each group. We also present the number of classes, actors, and video clips along with their frame resolution.

Many of the existing datasets for human activity recognition were recorded in controlled environments, with participant actors performing specific actions. Furthermore, several datasets are not generic, but rather cover a specific set of activities, such as sports or simple actions, which are usually performed by one actor. However, these limitations

Table 2.2: Comparison of unimodal methods.

Type of method	**Pros**	**Cons**
Space-time	- Localization of actions	- Sensitivity to noise and occlusions
	- 3D body representation	- Recognizing complex activities may be tricky
	- Good representation of low-level features	- Feature sparsity leads to low repeatability
	- Detailed analysis of human movements	- Gap between low-level features and high-level events
	- Unsupervised learning	- Human body detection is often a prerequisite
Stochastic	- Complex activity recognition	- Learning and inference may be difficult
	- Modeling of human interactions	- Learning a large number of parameters
	- Recognition from very short clips	- Label bias problem
	- Partial occlusion, background clutter and camera motion handling	- Prone to overfitting
	- High generalization ability	- Approximate solutions
	- Non-periodic activity recognition	- Large number of training data required
Rule-based	- High-level representation of human actions	- Decomposition of complex activities into smaller tasks
	- Sequential activity recognition	- Only atomic actions are recognized
	- Context-free grammar classification	- Rule/attribute generation is difficult
	- Knowledge transfer between actions	- Problems with long video sequences
	- Learning of multiple tasks simultaneously	
Shape-based	- 2D and 3D body representation	- Large number of degrees of freedom
	- Independent modeling of human body parts	- Skeleton tracking inaccuracies
	- Recognition from still images	- View-point and self occlusions dependent
	- Upper body action recognition	- Sensitivity to illumination changes and human clothing
	- Existence of low cost devices for pose estimation	- Difficulties in mapping image feature space to pose space

constitute an unrealistic scenario that does not cover real-world situations and does not address the specifications for an ideal human activity dataset as presented earlier. Nevertheless, several activity recognition datasets that take into account these requirements have been proposed.

Several existing datasets have reached their expected life cycle (i.e., methods on Weizmann and KTH datasets achieved 100% recognition rate). These datasets were captured in controlled environments and the performed actions were obtained from a frontal view camera. The non-complex backgrounds and the non-intra-class variations in human movements make these datasets non-applicable for real world applications. However, these datasets still remain popular for human activity classification, as they provide a good evaluation criterion for many new methods. A significant problem in constructing a proper human activity recognition dataset is the annotation of each action, which is generally performed manually by the user, making the task biased.

Understanding human activities is a part of interpersonal relationships. Humans have

Table 2.3: Comparison of multimodal methods.

Type of method	**Pros**	**Cons**
Affective	- Association of human emotions and actions - Better understanding of human activities - Complex activity recognition - Incorporation of well known classification models	- Affective data annotation is difficult - Problems in handling continuous actions - Dimensionality of the different modalities - Gap between low-level features and high-level concepts
Behavioral	- Personalized action recognition - Improve human-computer interaction - Complex activity recognition - Recognizes human interactions - Psychological attributes improve recognition	- Emotional attribute specification is difficult - Mainly frontal view emotion recognition - Complex classification models - Proper feature selection is difficult - Visual feature descriptors cannot capture human emotions - Dimensionality of the different modalities
Social networking	- Recognizes social human interactions - Easy access to data though social platforms - Reliable recognition of human-to-human or human-to-object interactions - Abnormal activity recognition	- Limited by the number of interacting persons - Dimensionality of the different modalities - Decomposition of complex actions into smaller tasks is necessary - Difficulties in crowded scene modeling due to occlusions

the ability to understand another human's actions by interpreting stimuli from the surroundings. On the other hand, machines need a learning phase to be able to perform this operation. Thus, some basic questions arise about a human activity classification system:

1. How to determine whether a human activity classification system provides the best performance?

2. In which cases is the system prone to errors when classifying a human activity?

3. In what level can the system reach the human ability of recognizing a human activity?

4. Are the success rates of the system adequate for inferring safe conclusions?

It is necessary for the system to be fully automated. To achieve this, all stages of human activity modeling and analysis are to be performed automatically, namely: (i) human activity detection and localization, where the challenge is to detect and localize a human activity in the scene. Background subtraction [12] and human tracking [14] are usually used as part of this process; (ii) Human activity modeling (e.g., feature extraction [97]) is the step of extracting the necessary information that will help in the recognition step; and (iii) human activity classification is the step where a probe video sequence is classified in one of the classes of the activities that have been defined before building the system.

Table 2.4: Human activity recognition datasets.

Dataset name and category	Year	# Classes	# Actors	# Videos	Resolution
Single action recognition					
KTH [2]	2004	6	25	2,391	160×120
Weizman [1]	2005	10	9	90	180×144
UCF Sports [3]	2008	9		200	720×480
MuHAVi [276]	2010	17	14		720×576
UCF50 [277]	2013	50		6,676	
UCF101 [278]	2012	101		13,320	320×240
Movie					
UCF YouTube [4]	2009	11		> 1.100	720×480
Hollywood2 [249]	2009	12		3,669	
HMDB51 [279]	2011	51		6,849	320×240
TVHI [6]	2012	4	20	300	320×240
Surveillance					
PETS 2004 (CAVIAR) [280]	2004	6		28	384×288
PETS 2007 [281]	2007	3		7	768×576
VIRAT [282]	2011	23		17	1920×1080
Pose					
TUM Kitchen [283]	2009	10	4	20	324×288
Two-person interaction [7]	2012	8	7	≈ 300	640×480
MSRC-12 Kinect gesture [284]	2012	12	30	594	
J-HMDB [285]	2013	21	1	928	240×320
UPCV action [204]	2014	10	20	≈ 200	
Daily living					
URADL [104]	2009	17	5	150	1280×720
ADL [17]	2012	18	20	≈ 10 hours	1280×960
MPII Cooking [286]	2012	65	12	44	1624×1224
Breakfast [145]	2014	10	52	≈ 77 hours	320×240
Social networking					
CCV [287]	2001	20		9,317	
FPSI [242]	2012	6	8	≈ 42 hours	1280×720
Broadcast field hockey [248]	2012	11		58	
USAA [8]	2012	8		≈ 200	
Sports-1M [262]	2014	487		1M	
ActivityNet [288]	2015	203		27,801	1280×720
WWW Crowd [260]	2015	94		10,000	640×360
Behavior					
BEHAVE [289]	2007	8		321	640×480
Canal9 [290]	2009	2	190	≈ 42 hours	720×576
USC Creative IT [291]	2010	50	16	100	
Parliament [5]	2014	3	20	228	320×240

In addition, the system should work regardless of any external factors. This means that the system should perform robustly despite changes in lighting, pose variations or partially occluded human bodies, and background clutter. Also, the number as well as the type of

human activity classes to be recognized is an important factor that plays a crucial role in the robustness of the system. The requirements of an ideal human activity classification system should cover several topics, including automatic human activity classification and localization, lighting and pose variations (e.g., multi-view recognition), partially occluded human bodies, and background clutter. Also, all possible activities should be detected during the recognition process, the recognition accuracy should be independent from the number of activity classes, and the activity identification process should be performed in real time and provide a high success rate and low false positive rate.

Chapter 3

Matching Mixtures of Trajectories for Human Action Recognition

3.1 Introduction

3.2 Action Representation and Recognition

3.3 Experimental Results

3.4 Conclusion

3.1 Introduction

In this chapter, we address the problem of human action recognition by representing an action with a set of clustered motion curves. Motion curves are generated by optical flow features which are then clustered using a different Gaussian mixture [25] for each distinct action. The optical flow curves of a probe sequence are also clustered using a Gaussian mixture model (GMM) and they are matched to the learned curves using a similarity function [292] relying on the longest common subsequence (LCSS) between curves and the canonical time warping (CTW) [293]. Linear [25] and non linear [294] dimensionality reduction methods may also be employed in order to remove outliers from the motion curves and reduce their lengths. The motion curve of a new probe video is projected onto its own subspace by a projection matrix specified by that video, and then the action label of the closest projection is selected according to the learned feature vectors as the identity of the probe sequence. The LCSS is robust to noise and provides an intuitive notion of similarity between curves. Since different actors perform the same action in different manners and at different speeds, an advantage of the LCSS similarity is that it can handle

with motion curves of varied lengths. On the other hand, CTW, which is based on the dynamic time warping [169], allows the spatio-temporal alignment between two human motion sequences. A preliminary version of this work was presented in [48]. One of the main contributions of this work, is that the training sequences do not need to have the same length. When a new probe sequence comes, it is matched against all the training sequences using the LCSS similarity measure. This measure provides a similarity between motion curves without enforcing one-to-one matching. An optimal matching is performed using dynamic programming, which detects similar pairs of curve segments [292].

However, training an action recognition system with only the knowledge of the motion of the current subject it is on its own a challenging task. The main problem is how we can ensure the continuity of the curves along time as an action occurs uniformly or non-uniformly within a video sequence. Unlike other approaches [62, 63], which use snippets of motion trajectories, our approach uses the full length of motion curves by tracking the optical flow features. Another question concerns the optimal model that one should adopt for recognizing human actions with high accuracy. This is accomplished by a statistical measure based on the data likelihood. The different lengths of the video sequences and therefore the respective lengths of the motion curves is another problem that is addressed. The large variance between benchmark datasets shows how the algorithm may be generalized. All these problems are discussed here and proper solutions are proposed. To this end, we have conducted experiments on several datasets [2, 3, 4] that would help us to understand how human activity recognition works.

Concatenating of optical flow features along time allows us to collect time series that preserve their continuity along time. It is true that correspondence is missing. However, this is the main assumption in many works [65, 66, 69]. If data association were used the resulting feature curves would have short duration and would be incomplete, as the features disappear and reappear due to occlusion, illumination, viewpoint changes and noise. In that case, a combination of sparse approach of clustering curves with variant lengths and tracking approaches should be used [295, 296]. This is not the central idea in this chapter, as the nature of the feature curves drastically changes.

3.2 Action Representation and Recognition

Our goal is to analyze and interpret different classes of actions to build a model for human activity categorization. Given a collection of figure-centric sequences, we represent motion templates using optical flow [297] at each frame. Assuming that a bounding box can be automatically obtained from the image data, we define a rectangle region of interest (ROI) around the human. A brief overview of our approach is depicted in Figure 3.1. In the training mode, we assume that the video sequences contain only one actor performing only one action per frame. However, in the recognition mode, we allow more than one action per video frame. The optical flow vectors as well as the motion descriptors [65] for each sequence are computed. These motion descriptors are collected together to construct

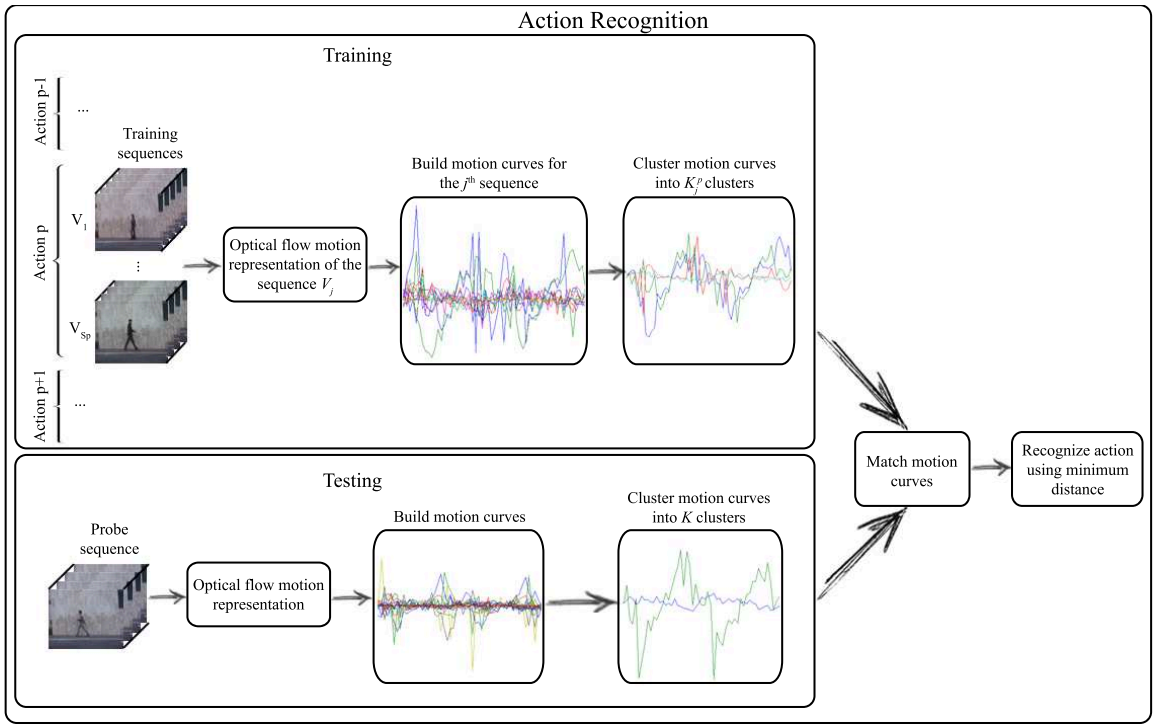


Figure 3.1: Overview of our approach.

motion curves, which are clustered using a mixture model to describe a unique action. Then, the motion curves are clustered and each action is modeled by a set of clustered motion curves. Action recognition is performed by matching the clusters of motion curves of the probe sequence and the clustered curves in each training sequence.

3.2.1 Motion Representation

The proposed approach employs optical flow features [297]. These motion descriptors are commonly used in many recognition problems and they are shown to be quite reliable despite the existence of noisy features. Within a figure-centric scene, any human motion may be decomposed to the motion of different body parts (e.g., head and limbs). We can easily localize the motion by computing the optical flow vectors for the regions around the human torso.

Following the work of Efros *et al.* [65], we compute the motion descriptor for the ROI as a four-dimensional vector $\mathbf{F}_i = \left(F^+_{x_i}, F^-_{x_i}, F^+_{y_i}, F^-_{y_i}\right) \in \mathbb{R}^4$, where $i = 1, ..., N$, with N being the number of pixels in the ROI. Also, the matrix $\mathbf{F}$ refers to the blurred, motion compensated optical flow. We compute the optical flow $\mathbf{F}$, which has two components, the horizontal $\mathbf{F}_x$, and the vertical $\mathbf{F}_y$, at each pixel. It is worth noting that the horizontal and vertical components of the optical flow $\mathbf{F}_x$ and $\mathbf{F}_y$ are half-wave rectified into four non-negative channels F^+_x,F^-_x,F^+_y,F^-_y, so that $\mathbf{F}_x = F^+_x - F^-_x$ and $\mathbf{F}_y = F^+_y - F^-_y$. In the general case, optical flow is suffering from noisy measurements and analyzing data under these circumstances will lead to unstable results. To handle any motion artifacts

due to camera movements, each half-wave motion compensated flow is blurred with a Gaussian kernel. In this way, the substantive motion information is preserved, while minor variations are discarded. Thus, any incorrectly computed flows are removed. Since all curves are considered normally distributed there is an intrinsic smoothing of the optical flow curves. Moreover, at a preprocessing step, we discard flows whose amplitude is over 20% of the standard deviation of the mean amplitude of all curves for each video.

3.2.2 Extraction of Motion Curves

A human action is represented by a set of primitive motion curves which are constructed directly from the optical flow motion descriptors. The main idea is to extract the salient features, which describe a relative motion from each frame and associate them with the corresponding feature in the next frame.

Consider T to be the number of image frames and C $=\{c_i(t)\}, t \in [0, T]$, is a set of motion curves for the set of pixels $i = 1, ..., N$ of the ROI. Each motion curve is described as a set of points corresponding to the optical flow vector extracted in the ROI. Specifically, we describe the motion at each pixel by the optical flow vector $\mathbf{F}_i = \left(F^+_{x_i}, F^-_{x_i}, F^+_{y_i}, F^-_{y_i}\right)$. A set of motion curves for a specific action is depicted in Figure 3.1. Given the set of motion descriptors for all frames, we construct the motion curves by following their optical flow components in consecutive frames. If there is no pixel displacement we consider a zero optical flow vector displacement for this pixel.

The set of motion curves describes completely the motion in the ROI. Once the motion curves are created, pixels and therefore curves that belong to the background are eliminated. We assume that the motion are normally distributed, thus, we keep flows whose values are inside 6 standard deviations of the amplitude distributions. In order to establish a correspondence between the motion curves and the actual motion, we perform clustering of the motion curves using a Gaussian mixture model. We estimate the characteristic motion which is represented by the mean trajectory of each cluster.

3.2.3 Motion Curves Clustering

A motion curve is considered to be a 2D time signal:

$$c_{ji}(t) = \left(F_{x_{ji}}(t), F_{y_{ji}}(t)\right), \;\; t \in [0, T], \tag{3.1}$$

where the index $i = 1, \ldots, N$ represents the i^{th} pixel, for the j^{th} video sequence in the training set. To efficiently learn human action categories, each action is represented by a GMM by clustering the motion curves in every sequence of the training set. The p^{th} action ($p = 1, ..., A$), in the j^{th} video sequence ($j = 1, \ldots, S_p$), is modeled by a set of K^p_j mean curves learned by a GMM. The likelihood of the i^{th} curve $c^p_{ji}(t)$ of the p^{th} action in the j^{th} video is given by:

$$p(c^p_{ji}; \pi^p_j, \mu^p_j, \Sigma^p_j) = \sum_{k=1}^{K^p_j} \pi^p_{jk} \mathcal{N}(c^p_{ji}(t); \mu^p_{jk}, \Sigma^p_{jk}), \;\; t \in [0, T], \tag{3.2}$$

where $\pi_j^p = \{\pi_{jk}^p\}_{k=1}^{K_j^p}$ are the mixing coefficients, $\mu_j^p = \{\mu_{jk}^p\}_{k=1}^{K_j^p}$ is the set of the mean curves and $\Sigma_j^p = \{\Sigma_{jk}^p\}_{k=1}^{K_j^p}$ is the set of covariance matrices. The covariance matrix in equation (3.2) is a diagonal $\Sigma_{jk}^p = diag(\sigma_{jk,1}^{2p}, \ldots, \sigma_{jk,T}^{2p})$. Therefore, the log-likelihood of the p^{th} action in the j^{th} video can be written as:

$$L(c_j^p) = \prod_{i=1}^{N_j^p} \ln \sum_{k=1}^{K_j^p} \pi_{jk}^p \mathcal{N}(c_{ji}^p(t); \mu_{jk}^p, \Sigma_{jk}^p), \;\; t \in [0, T]\,, \tag{3.3}$$

where N_j^p is the number of motion curves in the training set describing the p^{th} action in the j^{th} video.

The GMM is trained using the Expectation-Maximization (EM) algorithm [25], which provides a solution to the problem of estimating the model's parameters. The initialization of the EM algorithm is performed by the K-means algorithm. We have examined several configurations for the initialization of K-means and we decided to employ K-means with 50 different random initializations which were consistent and had no significant impact on the final classification. However, the number of mixture components should be determined. To select the number of the Gaussians K_j^p, for the j^{th} training video sequence, representing the p^{th} action, the Bayesian Information criterion (BIC) [25] is used:

$$BIC(c_j^p) = L(c_j^p(t)) - \frac{1}{2} M N_j^p, \;\; t \in [0, T]\,, \tag{3.4}$$

where M is the number of parameters of the GMM to be inferred. Thus, when EM converges the cluster labels of the motion curves are obtained. This is schematically depicted in Figure 3.1, where a set of motion curves, representing a certain action (e.g., p), in a video sequence (e.g., labeled by j) is clustered by a GMM into $K_j^p = 2$ curves for action representation. Note that a given action is generally represented by a varying number of mean curves as the BIC criterion may result in a different number of components in different sequences.

Apart from the BIC criterion, there are other techniques for determining the appropriateness of a model such as the Akaike Information Criterion (AIC) [25].

$$AIC(c_j^p) = L(c_j^p(t)) - M, \;\; t \in [0, T]\,, \tag{3.5}$$

where M is the number of parameters of the GMM to be inferred. BIC is independent of the prior, it can measure the efficiency of the parameterized model in terms of predicting the data and it penalizes the complexity of the model, where complexity refers to the number of parameters in the model. It is also approximately equal to the minimum description length criterion [25] but with negative sign, it can be used to choose the number of clusters according to the intrinsic complexity present in a particular dataset and it is closely related to other penalized likelihood criteria such as the AIC. BIC tends to select highly parsimonious models, while AIC tends to include more parameters [298, 299]. Complexity measures such as BIC and AIC have the virtue of being easy to evaluate, but can also give misleading results.

3.2.4 Matching of Motion Curves

Once a new probe video is presented, where we must recognize the action depicted, the optical flow is computed, motion curves are created and clustered, and they are compared with the learned mean curves of the training set. Recall that human actions are not uniform sequences in time, since different individuals perform the same action in different manner and at different speeds. This means that motion curves have varied lengths. An optimal matching may be performed using dynamic programming which detects similar pairs of curve segments. The longest common subsequence (LCSS) [292] is robust to noise and provides a similarity measure between motion curves since not all points need to be matched.

Let $\mu(t)$, $t \in [0, T]$ and $\nu(\tau)$, $\tau \in [0, T']$ be two curves of different lengths. Then, we define the affinity between the two curves as:

$$\alpha\left(\mu(t), \nu(\tau)\right) = \frac{LCSS\left(\mu(t), \nu(\tau)\right)}{\min(T, T')}\,, \tag{3.6}$$

where the $LCSS\left(\mu(t), \nu(\tau)\right)$ (Eq. (3.7)) indicates the quality of the matching between the curves $\mu(t)$ and $\nu(\tau)$ and measures the number of the matching points between two curves of different lengths.

$$LCSS\left(\mu(t), \nu(\tau)\right) =$$
$$\begin{cases} 0, & \text{if} \quad T = 0 \quad \text{or} \quad T' = 0\,, \\ 1 + LCSS\left(\mu(t)^{T_t - 1}, \nu(\tau)^{T'_\tau - 1}\right), & \text{if} \quad |\mu(t) - \nu(\tau)| < \varepsilon \quad \text{and} \quad |T - T'| < \delta\,, \\ \max\left\{LCSS\left(\mu(t)^{T_t - 1}, \nu(\tau)^{T'_\tau}\right),\ LCSS\left(\mu(t)^{T_t}, \nu(\tau)^{T'_\tau - 1}\right)\right\}, & \text{otherwise}\,. \end{cases} \tag{3.7}$$

Note that the LCSS is a modification of the edit distance [169] and its value is computed within a constant time window δ and a constant amplitude ε, that control the matching thresholds. The terms $\mu(t)^{T_t}$ and $\nu(\tau)^{T'_\tau}$ denote the number of curve points up to time t and τ, accordingly. The idea is to match segments of curves by performing time stretching so that segments that lie close to each other (their temporal coordinates are within δ) can be matched if their amplitudes differ at most by ε. A characteristic example of how two motion curves are matched is depicted in Figure 3.2.

When a probe video sequence is presented, its motion curves $z = \{z\}_{i=1}^{N}$ are clustered using GMMs of various numbers of components using the EM algorithm. The BIC criterion is employed to determine the optimal value of the number of Gaussians K, which represent the action in the probe sequence. Thus, we have a set of K mean curves ν_k, $k = 1, \ldots, K$ modeling the probe action, whose likelihood is given by:

$$L(z) = \prod_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(z_i; \nu_k, \Sigma_k)\,, \tag{3.8}$$

where Σ_k is the covariance matrix for the k^{th} component.

Recognition of the action present in the probe video sequence is performed by assigning the probe action to the action of the labeled sequence which is most similar. As both the

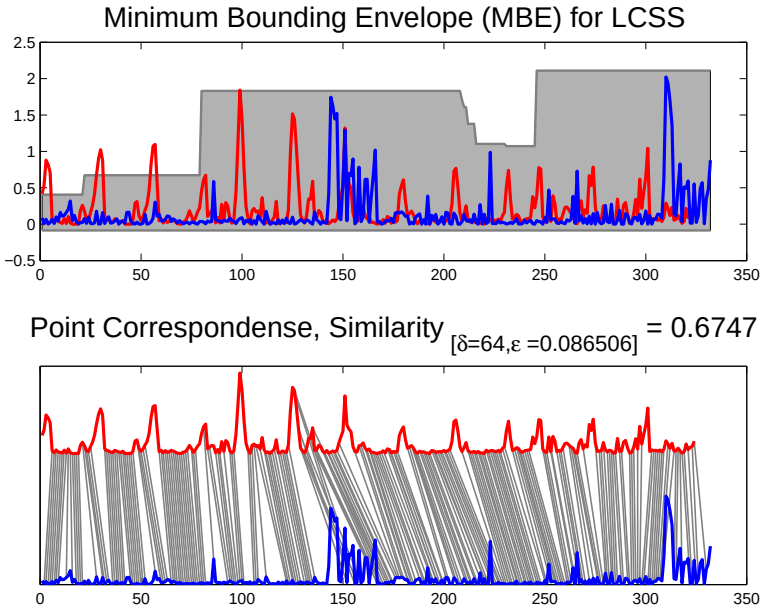


Figure 3.2: Depiction of the LCSS matching between two motions considering that they should be within $\delta = 64$ time steps in the horizontal axis and their amplitudes should differ at most by $\varepsilon = 0.086$.

probe sequence and the j^{th} labeled video sequence of the p^{th} action in the training set are represented by a number of mean curves $\nu = \{\nu_i\}_{i=1}^{K}$ and $\mu_j^p = \{\mu_{jk}^p\}_{k=1}^{K_j^p}$ respectively, the overall distance between them is computed by:

$$d(\mu_j^p, \nu) = \sum_{k=1}^{K_j^p} \sum_{\ell=1}^{K} \pi_{jk}^p \pi_\ell \left[1 - \alpha\left(\mu_{jk}^p(t), \nu_\ell(\tau)\right)\right] , \tag{3.9}$$

where π_{jk}^p and π_ℓ are the GMM mixing proportions for the labeled and probe sequence, respectively, that is $\sum_k \pi_{jk}^p = 1$ and $\sum_\ell \pi_\ell = 1$. The probe sequence ν is categorized with respect to its minimum distance from an already learned action:

$$[j^*, p^*] = \arg\min_{j,p} d(\mu_j^p, \nu) . \tag{3.10}$$

3.2.5 Canonical Time Warping

The canonical time warping (CTW) [293] solves the problem of spatio-temporal alignment of human motion between two time series. Based on dynamic time warping, the algorithm in [169] finds the temporal alignment of two subjects maximizing the spatial correlation between them. Given two time series $\mathcal{C}_1 = [c_1(0), \ldots, c_1(T)]$ and $\mathcal{C}_2 = [c_2(0), \ldots, c_2(T')]$ canonical time warping minimizes the following energy function:

$$J_{ctw}(\mathbf{W}_{\mathcal{C}_1}, \mathbf{W}_{\mathcal{C}_2}, \mathbf{V}_{\mathcal{C}_1}, \mathbf{V}_{\mathcal{C}_2}) = \|\mathbf{V}_{\mathcal{C}_1}^\top \mathcal{C}_1 \mathbf{W}_{\mathcal{C}_1}^\top - \mathbf{V}_{\mathcal{C}_2}^\top \mathcal{C}_2 \mathbf{W}_{\mathcal{C}_2}^\top\|_F^2 , \tag{3.11}$$

where $\mathbf{W}_{\mathcal{C}_1}$ and $\mathbf{W}_{\mathcal{C}2}$ are binary selection matrices that need to be inferred to align $\mathcal{C}_1$ and $\mathcal{C}_2$, and $\mathbf{V}_{\mathcal{C}_1}$, $\mathbf{V}_{\mathcal{C}_2}$ parameterize the spatial warping by projecting sequences into the same coordinate system.

3.2.6 Dimensionality Reduction

Dimensionality reduction methods [294] may be employed in order to reduce the dimension of the motion curves and to enforce them to be of equal length. In the experiments, Principal Complements Analysis (PCA) [300] was chosen as a simple linear method but any other non-linear technique [294] could also be applied. When PCA is employed the time ordering is suppressed and curves are then transformed into feature vectors. In that case, the Bhattacharyya distance [169] is (among others) an appropriate matching measure.

Let v_1 and v_2, be two feature vectors following Gaussian distributions, with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2, respectively. The Bhattacharyya distance has the form :

$$d_B(v_{1j}^p, v_2) = \frac{1}{8}(\mu_1 - \mu_2)^\top \left(\frac{\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2}{2}\right)^{-1} (\mu_1 - \mu_2) + \frac{1}{2}\ln\left(\frac{\frac{\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2}{2}}{2\sqrt{|\mathbf{\Sigma}_1||\mathbf{\Sigma}_2|}}\right) . \tag{3.12}$$

To perform the match, one can project a probe video feature vector $v_2 = \{v_{2i}\}_{i=1}^K$ onto its own subspace by a projection matrix specified to that video and assume the label that lies closer than all the training feature vectors $v_{1j}^p = \{v_{1jk}^p\}_{k=1}^{K_j^p}$. For Gaussian mixture models, we define the Bhattacharyya distance as:

$$d_{GMM}(v_{1j}^p, v_2) = \sum_{k=1}^{K_j^p} \sum_{\ell=1}^{K} \pi_{jk}^p \pi_\ell d_B(v_{1jk}^p, v_{2\ell}) \,, \tag{3.13}$$

where π_{jk}^p and π_ℓ are the GMM mixing proportions for the labeled and probe sequence, respectively. This is common in GMM modeling [301]. The probe feature vector v_2 is categorized with respect to the minimum distance from an already learned action:

$$[j^*, p^*] = \arg\min_{j,p} d_{GMM}(v_{1j}^p, v_2) \,. \tag{3.14}$$

The overall approach for learning an action and categorizing a probe are summarized in Algorithm 1 and Algorithm 2, respectively. The steps inside the parenthesis indicate the extra steps when PCA is employed.

3.3 Experimental Results

In what follows, we refer to our mixtures of curves action recognition method by the acronym TMAR. We evaluated the proposed method on action recognition by conducting a set of experiments over publicly available datasets.

3.3.1 Evaluation over the Weizmann Dataset

First, we applied the algorithm to the Weizmann human action dataset [1]. The Weizmann dataset is a collection of 90 low-resolution videos, which consists of 10 different

Algorithm 1 Action learning

Input: Training video sequences.

Output: GMMs summarizing each action in each sequence.

1: **for** each action **do**
2: **for** each video sequence representing the action **do**
3: Compute the optical flow at each pixel and generate half-wave rectified features.
4: Construct the motion curves by concatenating the optical flow features.
5: (Perform dimensionality reduction of the motion curves.)
6: Cluster the motion curves by training GMMs with varying number of components and select the model maximizing the BIC criterion.
7: **end for**
8: **end for**

Algorithm 2 Action categorization

Input: A probe video sequence to be categorized and the GMMs summarizing the actions in the training sequences.

Output: Action label.

1: Compute the optical flow at each pixel of the probe sequence and generate half-wave rectified features.
2: Construct the motion curves by concatenating the optical flow features.
3: (Project the motion curves onto their own subspace by a projection matrix.)
4: Cluster the motion curves by training GMMs with varying number of components and select the model maximizing the BIC criterion.
5: Compute the distances between the GMM of the probe sequence and each GMM of the learnt actions.
6: Classify the probe sequence using a nearest neighbor classifier.

actions (i.e., run, walk, skip, jumping jack, jump forward, jump in place, gallop sideways, wave with two hands, wave with one hand, and bend), performed by nine different people. The videos were acquired with a static camera and contain uncluttered background. Nevertheless, the dataset provides a good evaluation context for testing the performance of the proposed algorithm, due to the periodicity of the actions. Figure 3.3 illustrates some sample frames from the Weizmann dataset.

To test the proposed method on action recognition we adopted the leave-one-out scheme. We learned the model parameters from the videos of eight subjects and tested the recognition results on the remaining video sequences. The procedure was repeated for all sets of video sequences and the final result is the average of the individual results. The optimal number of mixture components K_j^p for the j^{th} video sequence, $j = 1, \ldots, S_p$ of the p^{th} action $p = 1, \ldots, A$ is found by employing the BIC criterion. The value of BIC was computed for $K_j^p = 1$ to the square root of the maximum number of motion curves.

As shown in Table 3.1, the average correct classification of the algorithm on this

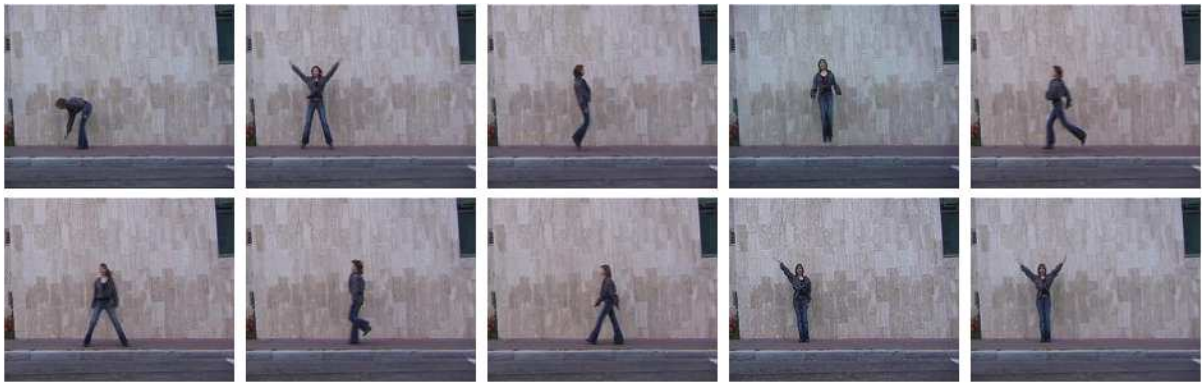


Figure 3.3: Sample frames from video sequences of the Weizmann dataset [1].

Table 3.1: Recognition accuracy over the Weizmann dataset.

Method	**Year**	**Accuracy (%)**
Blank *et al.* [1]	2005	100.0
Chaudhry *et al.* [72]	2009	95.7
Fathi and Mori [66]	2008	100.0
Jhuang *et al.* [67]	2007	98.8
Lin *et al.* [73]	2009	100.0
Niebles *et al.* [68]	2008	90.0
Seo and Milanfar [114]	2011	97.5
TMAR(LCSS-BIC)	2013	98.8
TMAR(CTW-BIC)	2013	92.2
TMAR(PCA-BIC)	2013	100.0

dataset is 98.8%, while it reaches 100% when the proposed method with PCA is utilized. However, the average correct classification falls to 92.2%, when the CTW is utilized. All motion curves are reduced to a length that explains the 90% of the eigenvalue sum, which results in a reduced curve length of 50 time instances with respect to the original 3.000 time instances. Note that better results are achieved with respect to four out of seven state-of-the-art methods for the standard method, whereas for the TMAR(PCA) the highest performance on this dataset is achieved. The proposed method provided only one erroneous categorization as one *jump-in-place* (pjump) action was incorrectly categorized as *run*. It appears that in this case the number of Gaussian components K_j^p computed by the BIC criterion was not optimal. Figure 3.4 depicts the confusion matrices for the TMAR(LCSS), TMAR(CTW) and TMAR(PCA) approaches.

More specifically, for the proposed method, when the LCSS metric is employed, for $K_j^p = 1$, $K_j^p = 2$ and $K_j^p = 3$ recognition rates of 100% are attained and performance begins to decrease for $K_j^p \geq 4$. This is not surprising since the majority of the mixture components provided by the BIC criterion is equal to two. In the case where CTW

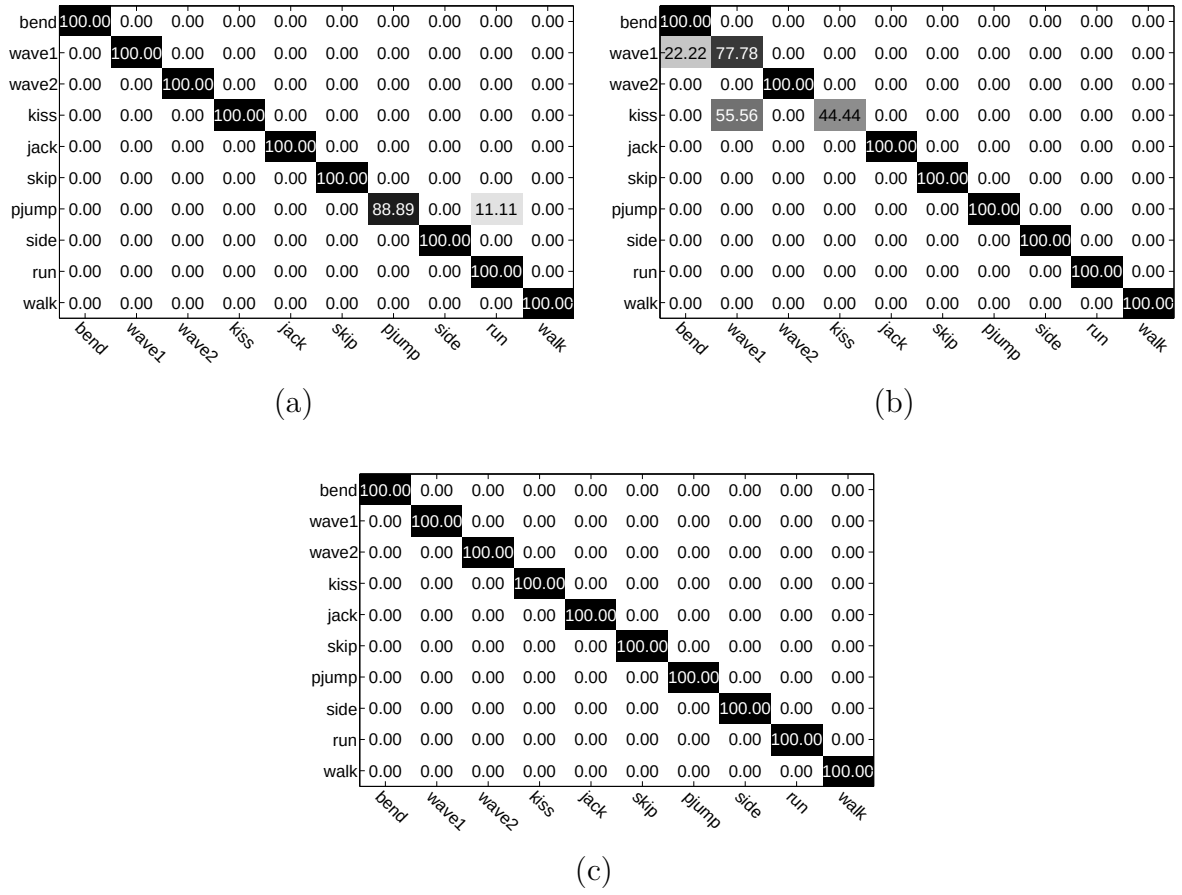


Figure 3.4: Confusion matrices of the classification results for the Weizmann dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion.

alignment is employed, the average recognition accuracy begins to fall for $K_j^p \geq 2$. When PCA is employed the recognition is perfect and begins to degrease for $K_j^p \geq 6$. In Figure 3.5, the recognition accuracy for this dataset with respect to the number of Gaussian components is depicted.

According to Tables 3.2 and 3.3, TMAR(LCSS-BIC) rejects the null hypothesis for five out of the seven cases. TMAR(CTW-BIC) rejects the null hypothesis for three out of seven cases but fails to reject the null hypothesis for the rest. In contrary to the previous approaches, TMAR(PCA-BIC) rejects the null hypothesis in all cases and is considered to be statistical significant.

3.3.2 Evaluation over the KTH Dataset

We also applied the proposed algorithm to the KTH dataset [2]. This dataset consists of 2.391 sequences and contains six types of human actions such as walking, jogging, running,

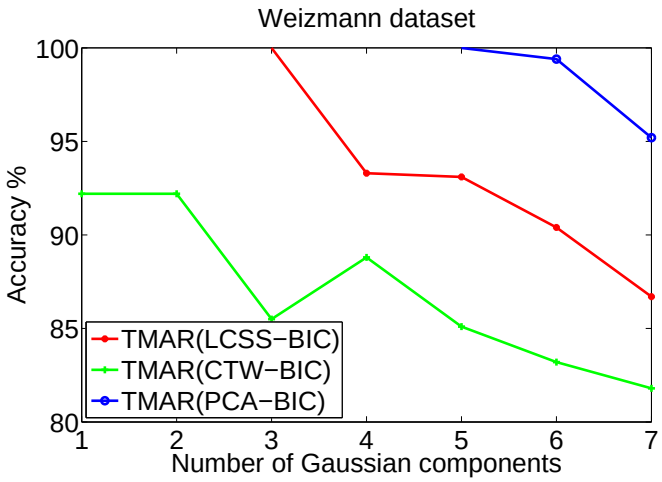


Figure 3.5: The recognition accuracy with respect to the number of Gaussian components for the Weizmann dataset.

Table 3.2: p-values for measuring the statistical significance of the proposed methods for the Weizmann dataset. The null hypothesis appears in the first column of the table.

Null hypothesis	**TMAR(LCSS-BIC)**	**TMAR(CTW-BIC)**	**TMAR(PCA-BIC)**
Blank *et al.* [1]	0.0283	0.0955	0.0097
Chaudhry *et al.* [72]	0.0097	0.0204	0.0254
Fathi and Mori [66]	0.0283	0.0955	0.0387
Jhuang *et al.* [67]	0.0723	0.0591	0.0455
Lin *et al.* [73]	0.0283	0.0955	0.0438
Niebles *et al.* [68]	0.0001	0.0450	0.0415
Seo and Milanfar [114]	0.1246	0.0393	0.0294

Table 3.3: Statistical measurements of the recognition results for each of the proposed approaches for the Weizmann dataset.

	mean	**median**	**std**	**min**	**max**
TMAR(LCSS-BIC)	98.9	100.0	3.5	88.8	100.0
TMAR(CTW-BIC)	92.2	100.0	18.2	44.4	100.0
TMAR(PCA-BIC)	100.0	100.0	0.0	100.0	100.0

boxing, hand waving, and hand clapping. These actions are repeatedly performed by 25 different people in four different environments: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors (s4). The video sequences were acquired using a static camera and include a uniform background. The average length of the video sequences is four seconds, while they were downsampled to a spatial resolution of 160×120 pixels. Figure 3.6 depicts sample snapshots from the KTH dataset.

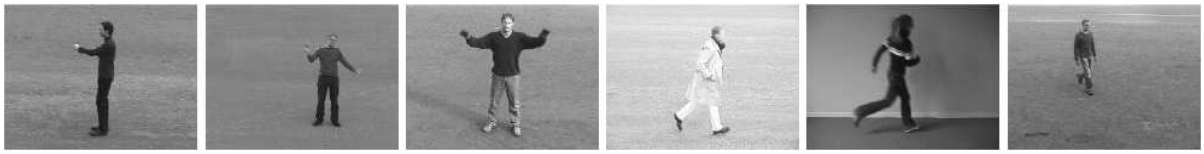


Figure 3.6: Sample frames from video sequences of the KTH dataset [2].

Table 3.4: Recognition results over the KTH dataset.

Method	**Year**	**Accuracy (%)**
Schuldt *et al.* [2]	2004	71.7
Jhuang *et al.* [67]	2007	90.5
Fathi and Mori [66]	2008	90.5
Niebles *et al.* [68]	2008	83.3
Lin *et al.* [73]	2009	95.8
Seo and Milanfar [114]	2011	95.1
Wang *et al.* [69]	2011	94.2
Wu *et al.* [98]	2011	94.5
Le *et al.* [95]	2011	93.9
Yan and Luo [78]	2012	93.9
Sadanand and Corso [96]	2012	98.2
TMAR(LCSS-BIC)	2013	96.7
TMAR(CTW-BIC)	2013	93.8
TMAR(PCA-BIC)	2013	98.3

We tested the action recognition performance of the proposed method by using a leave-one-out cross validation approach. Accordingly, the model from the videos of 24 subjects was learned while the algorithm was tested on the remaining subjects and averaged the recognition results. The confusion matrices over the KTH dataset for this leave-one-out approach are shown in Figure 3.7. A recognition rate of 96.7% was achieved when only the BIC criterion was employed in conjunction with the LCSS metric, 93.8% when the CTW alignment is employed, and 98.3% using PCA.

In addition, a comparison of the proposed method with other state-of-the-art methods is reported in Table 3.4. Note that the TMAR approach provides the more accurate recognition rates. All motion curves are reduced to a length that explains the 90% of the eigenvalue sum, which results in a reduced feature vector length of 50 instances with respect to the original $3,000$ time instances.

In order to examine the behavior and the consistency of the method to the BIC criterion, we have also applied the algorithm without using BIC but having a predetermined number of Gaussian components for both the training and the test steps. Therefore, we

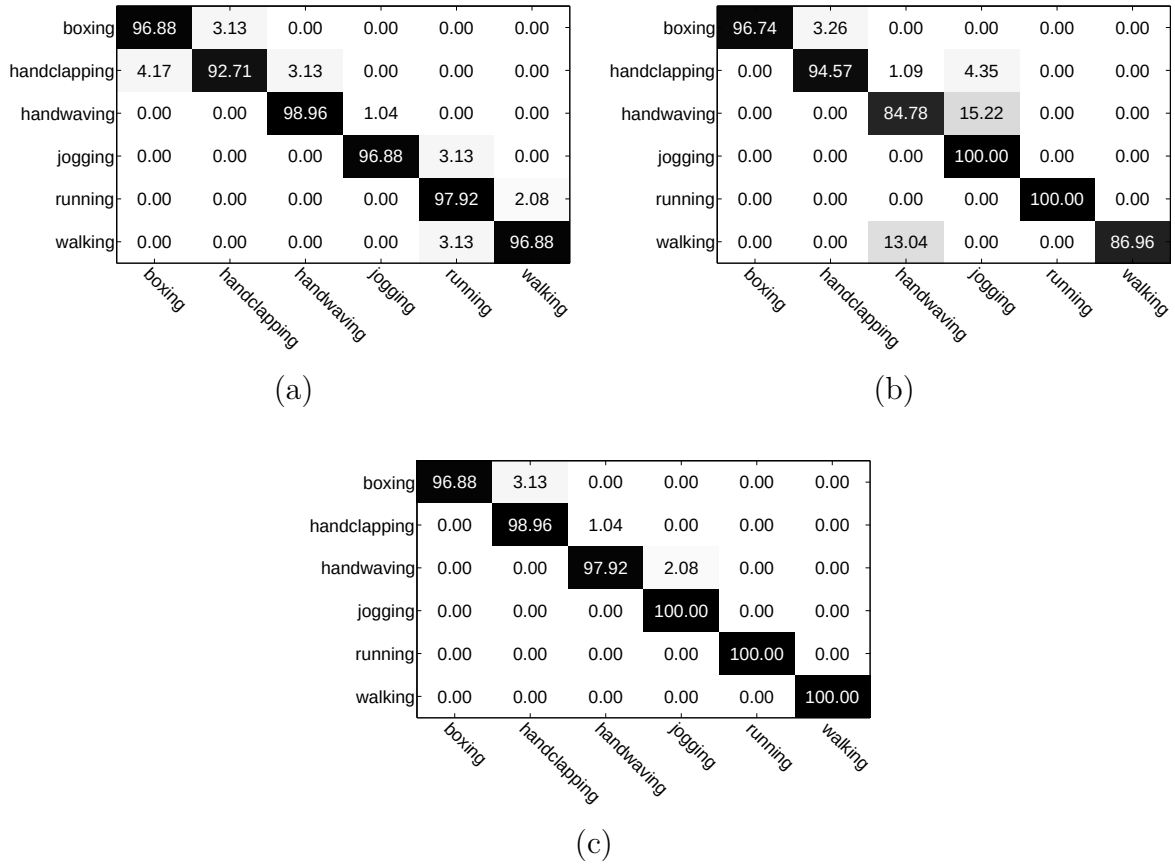


Figure 3.7: Confusion matrices of the classification results for the KTH dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion.

fixed the number of Gaussians K_j^p to values varying from one to the square root of the maximum number of the motion curves and executed the algorithm. The TMAR(LCSS) approach attains high action classification accuracy as the BIC criterion determines the optimal value of Gaussians K_j^p for this dataset. Figure 3.8 depicts the accuracy rate for the TMAR(LCSS), TMAR(CTW) and TMAR(PCA) approaches with respect to the number of mixture components. As the number of curves representing each action is relatively small (30–60 curves per action), a large number of Gaussian components may lead to model overfitting. As the number of Gaussians is $K_j^p \geq 3$ for the TMAR(LCSS), $K_j^p \geq 5$ for the TMAR(CTW) and $K_j^p \geq 4$ for the TMAR(PCA) the accuracy rate drastically falls. This fact indicates the dependency of the recognition accuracy over the number of Gaussian components as an action is represented by few motion curves.

In order to provide a statistical evidence of the recognition accuracy we present some statistical indices (Tables 3.5 and 3.6). The p-value is the probability of obtaining a statistical test at least as extreme as the one that was actually observed, assuming that the

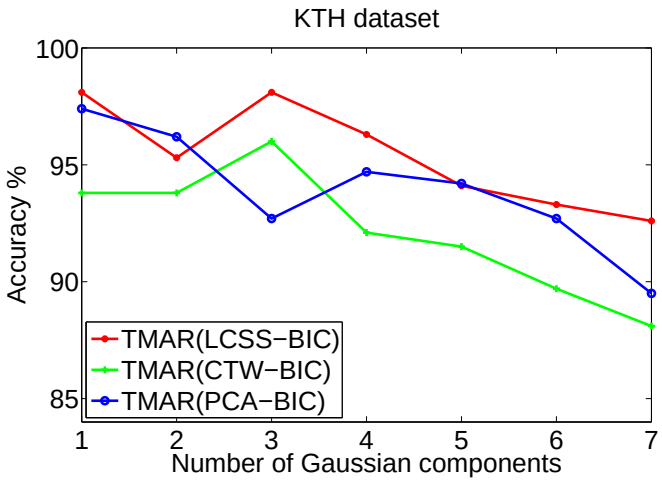


Figure 3.8: The recognition accuracy with respect to the number of Gaussian components for the KTH dataset.

Table 3.5: p-values for measuring the statistical significance of the proposed methods for the KTH dataset.

Method	**TMAR(LCSS-BIC)**	**TMAR(CTW-BIC)**	**TMAR(PCA-BIC)**
Schuldt *et al.* [2]	4.6142×10^{-7}	0.0002	3.9887×10^{-8}
Jhuang *et al.* [67]	4.2002×10^{-4}	0.1368	1.3577×10^{-5}
Fathi and Mori [66]	4.2002×10^{-4}	0.1368	1.3577×10^{-5}
Niebles *et al.* [68]	1.0195×10^{-5}	0.0056	6.3654×10^{-7}
Lin *et al.* [73]	0.1847	0.7551	0.0015
Seo and Milanfar [114]	0.0660	0.6757	5.9672×10^{-4}
Wang *et al.* [69]	0.0181	0.5563	2.2285×10^{-4}
Wu *et al.* [98]	0.0274	0.5977	3.0348×10^{-4}
Le *et al.* [95]	0.0121	0.5141	1.6643×10^{-4}
Yan and Luo [78]	0.0121	0.5141	1.6643×10^{-4}
Sadanand and Corso [96]	0.9340	0.9189	0.9389

null hypothesis is true. A small p-value ($p \leq 0.05$) indicates strong evidence against the null hypothesis, so we reject the null hypothesis. A large p-value ($p > 0.05$) indicates weak evidence against the null hypothesis, so we fail to reject the null hypothesis. Specifically, the null hypothesis was set to H_0: the recognition results of the state-of-the-art methods are better than the proposed and the alternative hypothesis is defined as H_α: the proposed method outperforms the state-of-the-art methods. In Table 3.5 and Table 3.6, statistical measurements for the KTH dataset are shown. TMAR(LCSS-BIC) and TMAR(PCA-BIC) reject the null hypothesis in the majority of the cases while, TMAR(CTW-BIC) rejects the null hypothesis in only two cases. Thus, the statistical significance meaning holds for TMAR(LCSS-BIC) and TMAR(PCA-BIC).

Table 3.6: Statistical measurements of the recognition results for each of the proposed approaches for the KTH dataset. All values are expressed in percentages.

	mean	**median**	**std**	**min**	**max**
TMAR(LCSS-BIC)	96.6	96.8	2.1	92.7	98.9
TMAR(CTW-BIC)	93.8	95.6	6.6	84.7	100.0
TMAR(PCA-BIC)	98.9	99.5	1.4	96.6	100.0

3.3.3 Evaluation over the UCF Sports Dataset

We have also applied our algorithm to the UCF Sports dataset [3]. This dataset consists of nine main actions such as diving, golf-swinging, kicking, lifting, horse riding, running, skating, swinging and walking. The dataset contains approximately 200 video sequences at a resolution of 720×480 pixels, which are captured in natural environment with a wide range of scenes and viewpoints. Figure 3.9 depicts some sample frames from the UCF Sports dataset.

Figure 3.9: Sample frames from video sequences of the UCF Sports dataset [3].

To test the proposed method on action recognition we also adopted the leave-one-out scheme. In Figure 3.10 are depicted the confusion matrices for the TMAR(LCSS), TMAR(CTW) and the TMAR(PCA) approaches. TMAR(LCSS) achieves 94.6% recognition accuracy with optimal number of components (BIC criterion) and 90.1% when the CTW alignment is employed. We also achieve the highest recognition accuracy of 95.1% when the proposed method uses PCA. In Figure 3.11, the dependency of the recognition accuracy with respect to the number of the Gaussian components is shown. For all three approaches as the number of components increases the recognition accuracy decreases, which may occur due to model overfitting. In the case where $K_j^p = 3$ all three approaches reach the highest peek of the graph. For $K_j^p \geq 4$ the recognition accuracy begins to decrease.

Table 3.7, shows the comparison between our TMAR approach, the baseline method using the BIC criterion in conjunction with the LCSS metric and the CTW alignment, the proposed method with PCA and previous approaches on the UCF Sports dataset. As it can be observed, the TMAR(PCA) approach preforms better than all the other methods,

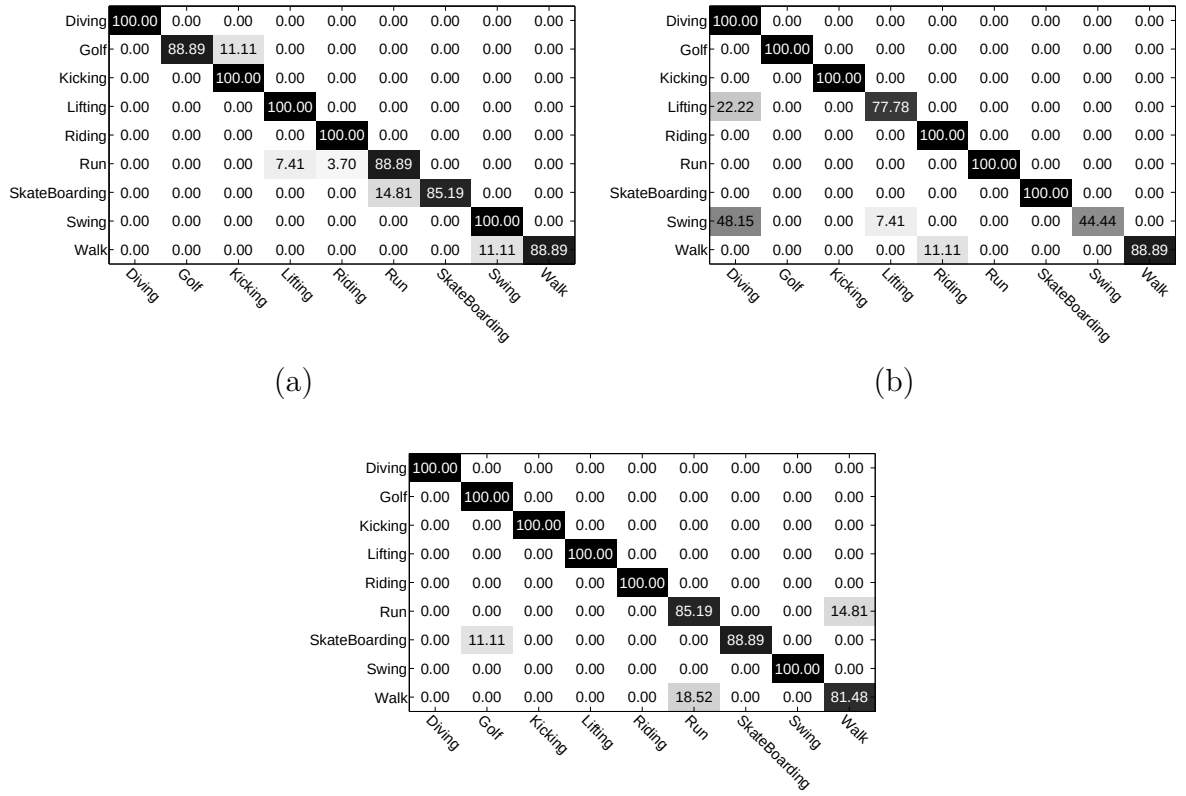


Figure 3.10: Confusion matrices of the classification results for the UCF Sports dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion.

while TMAR(LCSS) performs better for seven out of eight of the other methods. On the other hand, TMAR(CTW) has the less desirable performance as it outreaches four out of eight of the other methods on the same dataset.

Statistical evidence for the UCF Sports dataset is shown in Tables 3.8 and 3.9. TMAR(LCSS-BIC) and TMAR(PCA-BIC) appear to reject the null hypothesis for the majority of the cases, in contrary to the TMAR(CTW-BIC) which reject the null hypothesis only for the Rodriguez *et al.* [3] method. TMAR(LCSS-BIC) and TMAR(PCA-BIC) seem to be statistical significant while TMAR(CTW-BIC) is not.

3.3.4 Evaluation over the UCF YouTube Dataset

Finally, we have put our algorithm to test with the UCF YouTube dataset [4]. The UCF YouTube human action data set contains 11 action categories such as basketball shooting, biking, diving, golf swinging, horse riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This data set

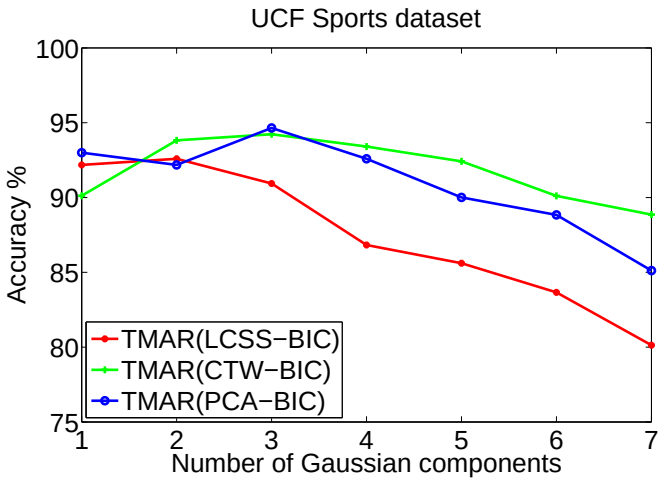


Figure 3.11: The recognition accuracy with respect to the number of Gaussian components for the UCF Sports dataset.

Table 3.7: Recognition results over the UCF Sport dataset.

Method	**Year**	**Accuracy (%)**
Rodriguez *et al.* [3]	2008	69.2
Kovaska and Grauman [115]	2010	87.3
Wang *et al.* [69]	2011	88.2
Wu *et al.* [98]	2011	91.3
Le *et al.* [95]	2011	86.5
Yan and Luo [78]	2012	90.7
Sadanand and Corso [96]	2012	95.0
TMAR(LCSS-BIC)	2013	94.6
TMAR(CTW-BIC)	2013	90.1
TMAR(PCA-BIC)	2013	95.1

includes actions with large variation in camera motion, object appearance and pose and scale. It also contains viewpoint and illumination changes, and spotty background. The video sequences are grouped into 25 groups of at least four actions each for each category, whereas the videos in the same group may share common characteristics such as similar background or actor. Representative frames of this data set are shown in Figure 3.12.

To assess our method we have used the leave-one-out cross validation scheme. In Figure 3.13 the confusion matrices for the TMAR(LCSS), TMAR(CTW) and TMAR(PCA) approaches are shown. We achieve a recognition rate of 91.7% when the LCSS metric is employed and having estimated the Gaussian components using the BIC criterion. We also achieve 91.3% when the CTW alignment is employed and 93.2% when using PCA. In Table 3.10, comparisons with other state-of-the-art methods for this dataset are reported.

Table 3.8: p-values for measuring the statistical significance of the proposed methods for the UCF Sports dataset.

Method	**TMAR(LCSS-BIC)**	**TMAR(CTW-BIC)**	**TMAR(PCA-BIC)**
Rodriguez *et al.* [3]	1.2614×10^{-6}	0.0052	3.9515×10^{-6}
Kovaska and Grauman [115]	0.0048	0.3336	0.0083
Wang *et al.* [69]	0.0090	0.3848	0.0142
Wu *et al.* [98]	0.0822	0.5735	0.0913
Le *et al.* [95]	0.0028	0.2909	0.0052
Yan and Luo [78]	0.0541	0.5369	0.0644
Sadanand and Corso [96]	0.5691	0.7714	0.4950

Table 3.9: Statistical measurements of the recognition results for each of the proposed approaches for the UCF Sports dataset. All values are expressed in percentages.

	mean	**median**	**std**	**min**	**max**
TMAR(LCSS-BIC)	94.6	100.0	6.5	85.1	100.0
TMAR(CTW-BIC)	90.1	100.0	18.8	44.4	100.0
TMAR(PCA-BIC)	95.1	100.0	7.7	81.4	100.0

Table 3.10: Recognition results over the UCF YouTube dataset.

Method	**Year**	**Accuracy (%)**
Liu *et al.* [4]	2009	71.2
Ikizler-Cinbis and Sclaroff [143]	2010	75.2
Le *et al.* [95]	2011	75.8
Wang *et al.* [132]	2011	84.2
TMAR(LCSS-BIC)	2013	91.7
TMAR(CTW-BIC)	2013	91.3
TMAR(PCA-BIC)	2013	93.2

As it can be seen, our algorithm achieves the highest recognition accuracy amongst all the others.

The performance of the proposed method with respect to the number of the Gaussian components is depicted in Figure 3.14. For TMAR(LCSS) the recognition accuracy begins to decrease for $K_j^p \geq 1$ and exhibits the worst performance than the other two approaches. The TMAR(CTW) approach decreases for $K_j^p \geq 2$ while TMAR(PCA) reaches its peak for $K_j^p = 4$ and then it begins to decrease. Note that the best approach tends to be attained by TMAR(PCA), which reaches a recognition accuracy of 91%.

Figure 3.12: Sample frames from video sequences of the UCF YouTube action dataset [4].

Table 3.11: p-values for measuring the statistical significance of the proposed methods for the UCF YouTube dataset.

Method	**TMAR(LCSS-BIC)**	**TMAR(CTW-BIC)**	**TMAR(PCA-BIC)**
Liu *et al.* [4]	2.1189×10^{-10}	4.0791×10^{-10}	9.6806×10^{-9}
Ikizler-Cinbis and Sclaroff [143]	1.7831×10^{-9}	3.5804×10^{-9}	6.6558×10^{-8}
Le *et al.* [95]	2.5596×10^{-9}	5.1805×10^{-9}	9.1874×10^{-8}
Wang *et al.* [132]	3.0663×10^{-6}	7.5786×10^{-6}	3.4593×10^{-5}

Table 3.12: Statistical measurements of the recognition results for each of the proposed approaches for the UCF YouTube dataset. All values are expressed in percentages.

	mean	**median**	**std**	**min**	**max**
TMAR(LCSS-BIC)	91.7	90.9	2.9	89.3	100.0
TMAR(CTW-BIC)	91.3	90.4	3.0	88.3	100.0
TMAR(PCA-BIC)	93.1	91.4	4.6	87.8	100.0

Table 3.11 and Table 3.12 present the same indices for the UCF YouTube dataset. All three proposed methods reject the null hypothesis for all the cases. In this case, the recognition results of the proposed methods for the UCF YouTube dataset appear to be statistically significant.

3.3.5 Parameter Estimation

In the recognition step, in our implementation of the LCSS the parameters δ and ϵ were optimized using 10-fold cross validation for all three datasets. These parameters need to be determined for each data set separately since each data set perform different types of actions. However, after we have determined the parameters no further action needs to be taken. To classify a new unknown sequence, we have already learned the parameters from the learning step and thus we are able to recognize the new action. For all the datasets,

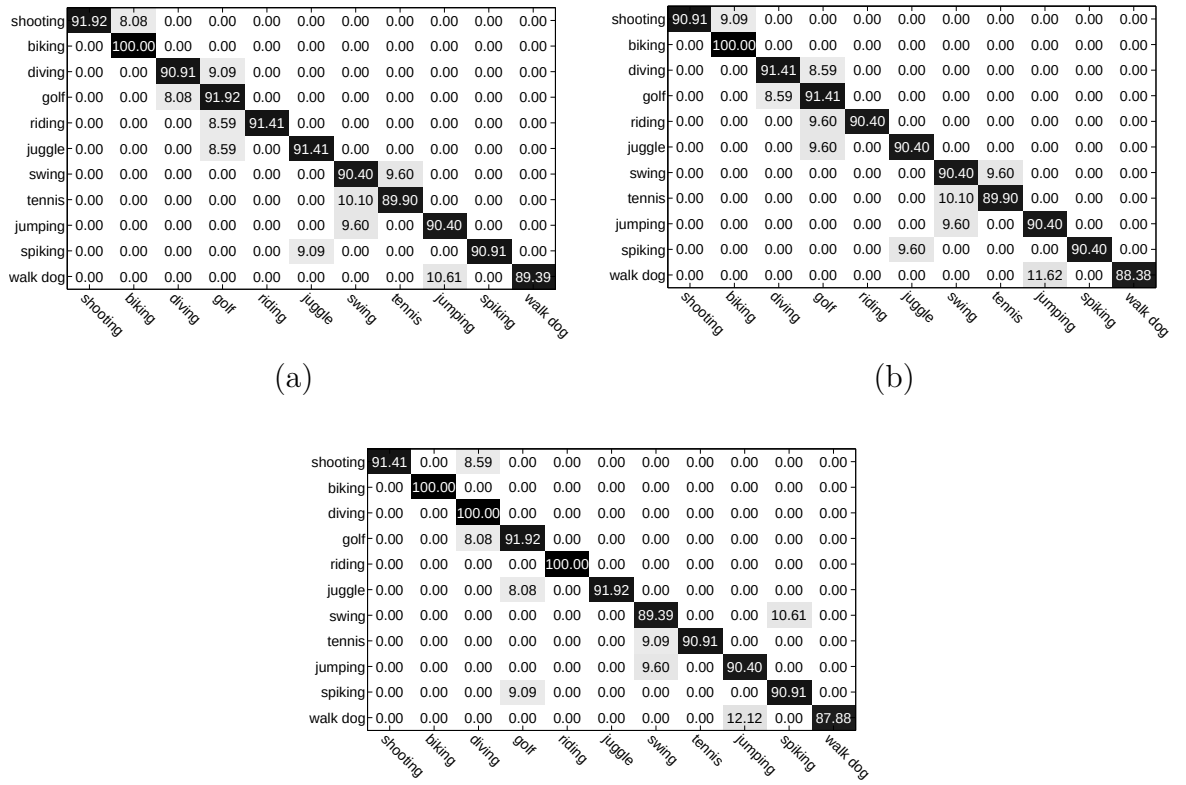


Figure 3.13: Confusion matrices of the classification results for the UCF YouTube dataset for (a) the proposed method denoted by TMAR(LCSS), (b) the the proposed method using the CTW alignment, denoted by TMAR(CTW), and (c) the proposed method using PCA, denoted by TMAR(PCA), for the estimation of the number of components using the BIC criterion.

Table 3.13, Table 3.14 and 3.15 show the optimal values per action as they have resulted after the cross validation process. Note that the values in Table 3.13 for both δ and ε are consistently small. However, the handclapping and walking actions have larger values for ε parameter than the other actions, which may be due to the large vertical movement of the subject between consecutive frames. On the other hand, the actions in the UCF Sport dataset holds large movements from one frame to the other for both horizontal and vertical axes, which is the main reason why the actions show large variances between the values of δ and ε (Table 3.14). Finally, the actions in the UCF YouTube dataset have a uniform distributed representation of the parameters δ and ε, since the parameter δ is determined as the 10% of the mean curves length for the most of the actions and the mean of the parameter ε is varies in the 15% of the standard deviation of the two curves to be compared.

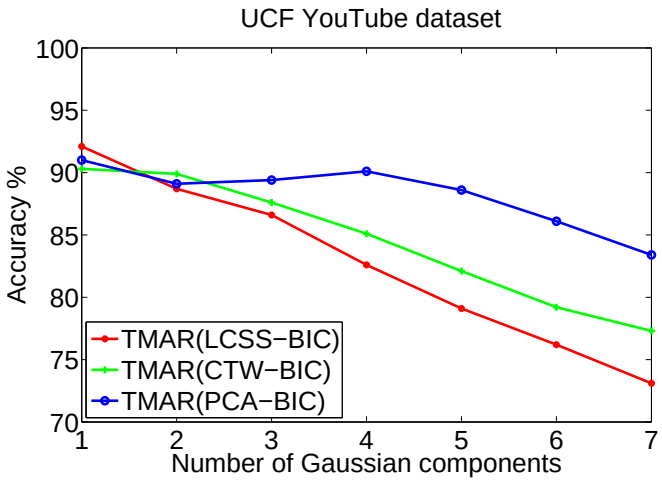


Figure 3.14: The recognition accuracy with respect to the number of Gaussian components for the UCF YouTube dataset.

Table 3.13: Parameters δ and ε for the KTH dataset estimated using cross validation.

Action	TMAR(LCSS)	
	$\delta(10^{-3})$	$\varepsilon(10^{-4})$
boxing	1	1
handclapping	10	100
handwaving	300	100000
jogging	5	3000
running	30	500
walking	1000	120000

3.3.6 Discussion

The average percentage of matched curves for the TMAR(LCSS) and TMAR(CTW) approach in the case where the BIC criterion is employed to determine the number of Gaussian components for all three datasets is depicted in Figure 3.15. As it can be observed, the TMAR(LCSS) method appears to match a larger part of curves for the same dataset than the TMAR(CTW) approach, which is the reason why TMAR(LCSS) performs better than TMAR(CTW).

In Figure 3.16, the execution times using the BIC criterion are depicted in order to determine the number of the Gaussian components, for all three cases, when using the LCSS metric, the CTW alignment and PCA, for all three datasets. For the Weizmann dataset, when PCA is used, the execution time drastically falls bellow one second per action. On the other hand, TMAR(LCSS) requires the highest execution time, which needs six seconds to recognize the action pjump. In the KTH dataset, TMAR(CTW) requires the highest execution time (needs nine seconds to recognize two out of six actions),

Table 3.14: Parameters δ and ε for the UCF Sports dataset estimated using cross validation.

Action	TMAR(LCSS)	
	δ	ε
diving	1	2.1
golf	2.01	6.1
kicking	10	15
lifting	11	10
riding	0.1	15
run	0.1	12
skateboarding	1.4	13
swing	0.6	20
walk	0.1	10

Table 3.15: Parameters δ and ε for the UCF Youtube dataset estimated using cross validation.

Action	TMAR(LCSS)	
	δ	ε
shooting	20	20
biking	10	10
diving	10	15
golf	20	10
riding	10	5
juggle	10	15
swing	10	5
tennis	10	10
jumping	10	5
spiking	10	30
walk dog	10	20

while TMAR(LCSS) takes less than six seconds for one action. Moreover, the use of PCA speeds up the execution time for recognizing a single action in all datasets since feature vectors of smaller lengths are being used. However, in UCF Sport dataset TMAR(LCSS) and TMAR(CTW) both have the same upper bound of eight seconds to recognize an action. Finally, in UCF YouTube dataset, the average execution time to recognize an action ranges from two to nine seconds when TMAR(LCSS) approach is used. In the case where TMAR(PCA) is used the upper bound to recognize an action is five seconds in UCF Sports and UCF YouTube datasets, while in in KTH is less than a second. This

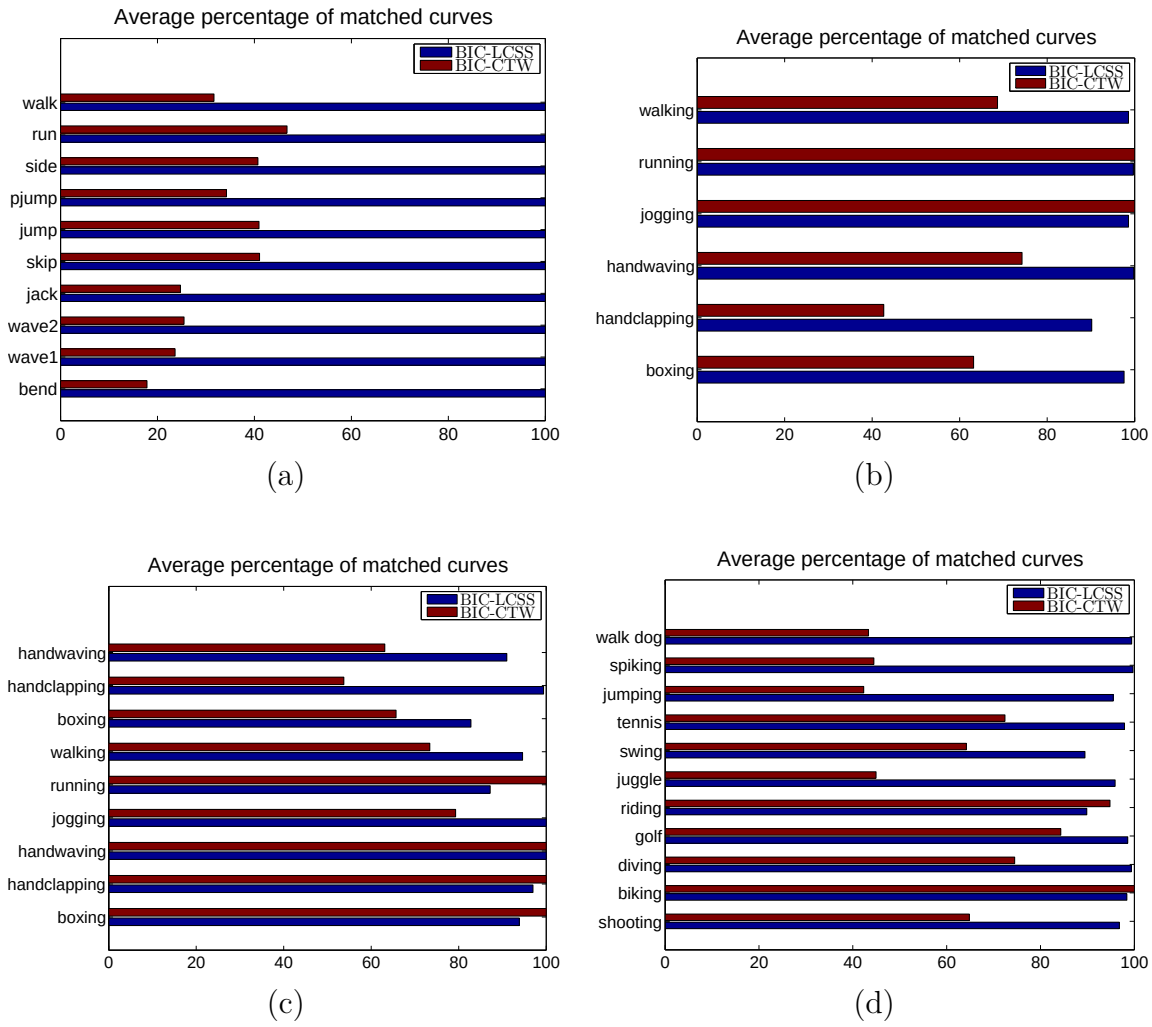


Figure 3.15: Average percentage of matched curves for TMAR(LCSS) and TMAR(CTW), when the BIC criterion is used, for (a) Weizman, (b) KTH, (c) UCF Sports and (d) UCF YouTube datasets, respectively.

makes the algorithm capable to adapt to any real video sequence and recognize an action really fast.

3.4 Conclusion

In this chapter, a human activity recognition method is proposed, where actions are represented by a set of motion curves generated by a probabilistic model. The performance of the extracted motion curves is interpreted by computing similarities between the motion curves, followed by a classification scheme. The large size of motion curves was reduced via PCA and after noise removal a reference database of feature vectors is obtained. Although a perfect recognition performance is accomplished with a fixed number of Gaussian mixtures, there are still some open issues in feature representation.

The obtained results showed that the use of PCA has a significant impact on the per-

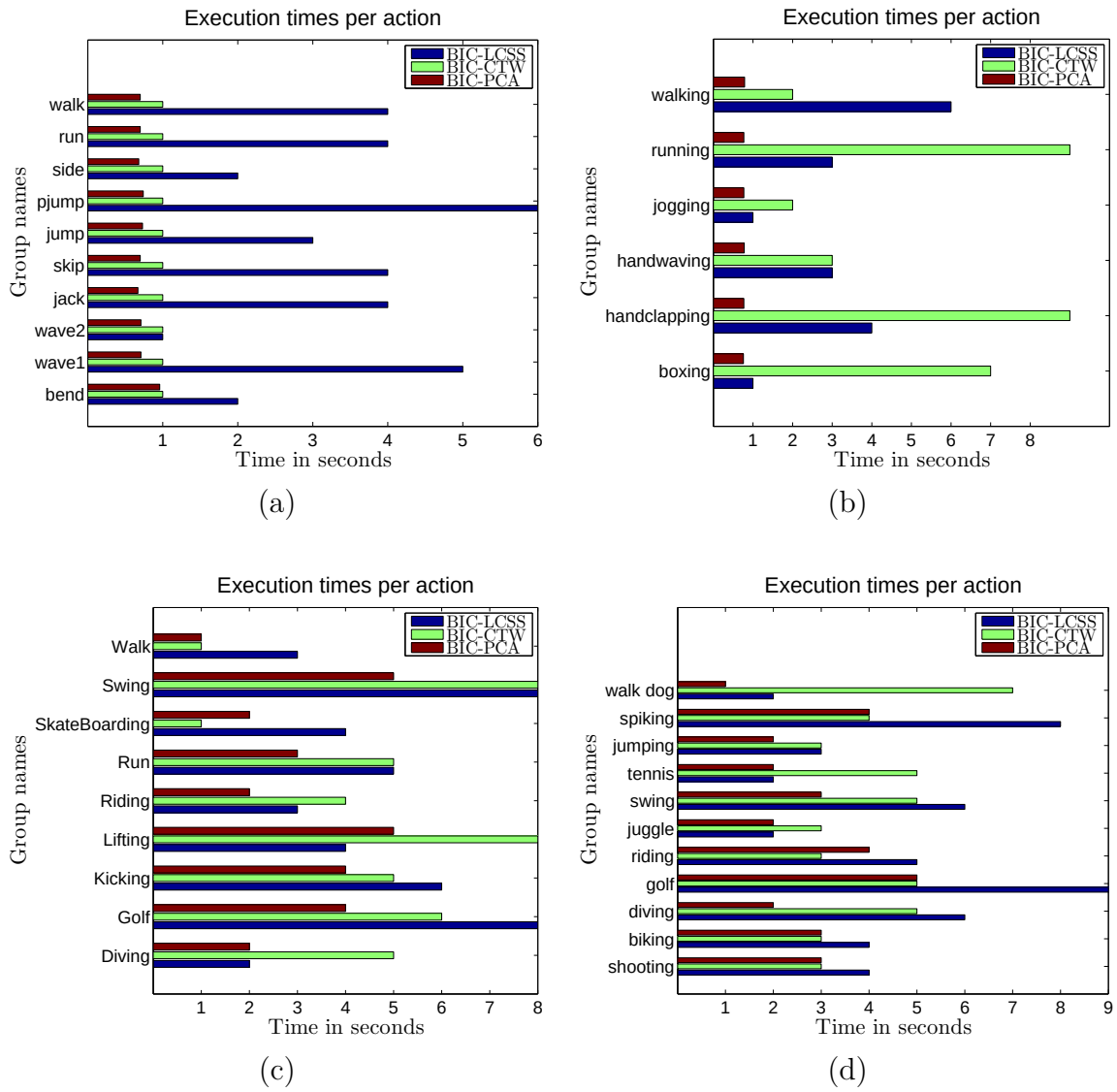


Figure 3.16: Execution times per action in seconds for TMAR(LCSS), TMAR(CTW) and TMAR(PCA), when the BIC criterion is used, for (a) KTH, (b) UCF Sports and (c) UCF YouTube datasets, respectively.

formance of the recognition process, as its use leads to further improvement of the recognition accuracy, while it significantly speeds up the behavior of the proposed algorithm. The optimal model was determined by using the BIC criterion. Finally, the presented algorithm is free of any constraints in the curves lengths. Although the proposed method yielded encouraging results in standard action recognition datasets, it is requirement of a challenging task of performing motion detection, background subtraction, and action recognition in natural and cluttered environments.

Chapter 4

Classifying Behavioral Attributes Using Conditional Random Fields

4.1 Introduction

4.2 Behavior Recognition Using Conditional Random Fields

4.3 Experimental Results

4.4 Conclusion

4.1 Introduction

In this chapter, we are interested in characterizing human activities as behavioral roles in video sequences. The main contribution of this work is twofold. First, we introduce a method for recognizing behavioral roles (i.e., friendly, aggressive and neutral) (Figure 4.1). These behavioral classes are similar, as the involved people perform similar body movements. Our goal is to recognize these behavioral states by building a model, which allows us to discriminate and correctly classify human behaviors. To solve this problem, we propose an approach based on conditional random fields (CRF) [26]. Motivated by the work of Domke [302], which takes into account both model and inference approximation methods to fit the parameters for several imaging problems, we develop a structured model for representing scenes of human activity and utilize a marginalization fitting for parameter learning. Secondly, to evaluate the model performance, we introduce a novel behavior dataset, which we call the *Parliament* dataset [5], along with the ground truth behavioral labels for the individuals in the video sequences. More specifically, we have collected 228 low-resolution video sequences (320×240, 25fps), depicting 20 different individuals speaking in the Greek parliament. Each video sequence is associated with a behavioral label: friendly, aggressive and neutral, depending on the intensity of the political speech and the specific individual's movements.

(a) (b) (c)

Figure 4.1: Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral

4.2 Behavior Recognition Using Conditional Random Fields

In this chapter, we present a supervised method for human behavior recognition. We assume that a set of training labels is provided and every video sequence is pre-processed to obtain a bounding box of the human in every frame and every person is associated with a behavioral label.

The model is general and can be applied to several behavior recognition datasets. Our method uses CRFs (Figure 4.2) as the probabilistic framework for modeling the behavior of a subject in a video. First, spatial local features are computed in every video frame capturing the roles associated with the bounding boxes. Then, a set of temporal context features are extracted capturing the relationship between the local features in time. Finally, the loopy belief propagation (LBP) [303] approximate method is applied to estimate the labels.

Let $\mathbf{y}_j^t \in \mathcal{Y}$ be the behavioral role label of the j^{th} person in a bounding box at frame t, where $\mathcal{R}$ is the set of possible behavioral role labels and $t \in [0, T]$ is the current frame. Let $\mathbf{x}_j^t$ represent the feature vector of the observed j^{th} pixel at frame t. Our goal is to assign each person a behavioral role by maximizing the posterior probability:

$$\mathbf{y} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}; \mathbf{w}) \,. \tag{4.1}$$

It is useful to note that our CRF model is a member of the exponential family defined as:

$$p(\mathbf{y}|\mathbf{x}; \mathbf{w}) = \exp\left(E(\mathbf{y}|\mathbf{x}; \mathbf{w}) - A(\mathbf{w})\right) \,, \tag{4.2}$$

where $\mathbf{w}$ is a vector of parameters, $E(\mathbf{y}|\mathbf{x})$ is a vector of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

$$A(\mathbf{w}) = \log \sum_{\mathbf{y}} \exp\left(E(\mathbf{y}|\mathbf{x}; \mathbf{w})\right) \,. \tag{4.3}$$

Different sufficient statistics $E(\mathbf{y}|\mathbf{x}; \mathbf{w})$ in (4.2) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$E(\mathbf{y}|\mathbf{x}; \mathbf{w}) = \sum_j \Psi_u(\mathbf{y}_j^t, \mathbf{x}_j^t; \mathbf{w}_1) + \sum_j \sum_{k \in \mathcal{N}_j} \Psi_p(\mathbf{y}_j^t, \mathbf{y}_k^{t+1}, \mathbf{x}_j^t, \mathbf{x}_k^{t+1}; \mathbf{w}_2) \,, \tag{4.4}$$

where $\mathcal{N}_j$ is the neighborhood system of the j^{th} person for every pixel in the bounding box. In our model temporal and spatial neighbors are considered. We use eight spatial and 18 temporal neighbors. The parameters w_1 and w_2 are the unary and the pairwise weights that need to be learned and $\Psi_u(\mathbf{y}_j^t, \mathbf{x}_j^t; \mathbf{w}_1)$, $\Psi_p(\mathbf{y}_j^t, \mathbf{y}_k^{t+1}, \mathbf{x}_j^t, \mathbf{x}_k^{t+1}; \mathbf{w}_2)$ are the unary and pairwise potentials, respectively.

Unary potential: This potential predicts the behavior label $\mathbf{y}_j^t$ of the j^{th} person in frame t indicating the dependence of the specific label on the location of the person. It may be expressed by:

$$\Psi_u(\mathbf{y}_j^t, \mathbf{x}_j^t; \mathbf{w}_1) = \sum_{\ell \in \mathcal{Y}} \sum_j \mathbf{w}_1^\top \mathbb{1}(\mathbf{y}_j^t = \ell) \psi_u(\mathbf{x}_j^t)\,, \tag{4.5}$$

where $\psi_u(\mathbf{x}_j^t)$ are the unary features and $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if the j^{th} person is associated with the ℓ^{th} label and 0 otherwise. The unary features are computed as a 36-dimensional vector of HoG3D values [241] for each bounding box. Then, a 64-dimensional spatio-temporal feature vector (STIP) [97] is computed, which captures the human motion between frames. The spatial relationship of each pixel in the bounding box and its 8×8 neighborhood is computed using a 16-dimensional Local Binary Pattern (LBP) feature vector [304]. The final unary features occur as a concatenation of the above features to a 116-dimensional vector.

Pairwise potential: This potential represents the interaction of a pair of behavioral labels in consecutive frames. We define the following function as the pairwise potential:

$$\Psi_p(\mathbf{y}_j^t, \mathbf{y}_k^{t+1}, \mathbf{x}_j^t, \mathbf{x}_k^{t+1}; \mathbf{w}_2) = \sum_{\substack{\ell \in \mathcal{Y}, \\ m \in \mathcal{Y}}} \sum_{\substack{j, \\ k \in \mathcal{N}_j}} \mathbf{w}_2^\top \mathbb{1}(\mathbf{y}_j^t = \ell) \mathbb{1}(\mathbf{y}_k^{t+1} = m) \psi_p(\mathbf{x}_j^t, \mathbf{x}_k^{t+1})\,, \tag{4.6}$$

where $\psi_p(\mathbf{x}_j^t, \mathbf{x}_k^{t+1})$ are the pairwise features. We compute a 4-dimensional spatio-temporal feature vector, which is the concatenation of the 2D velocity and acceleration of the j^{th}person along time. The acceleration features play a crucial role in the distinction between the behavioral classes, as different persons in different behavioral classes perform similar movements. In addition, the L_2 norm of the difference of the RGB values at frames t and $t+1$ is computed. We use eight spatial and 18 temporal neighbors creating an 18-dimensional feature vector. The final pairwise features are computed as the concatenation of the above features to a 22-dimensional vector.

4.2.1 Learning

To learn the model weights $\mathbf{w} = \{w_1, w_2\}$, we employ a labeled training set and seek to minimize:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{\mathbf{y}} L(\mathbf{y}, \mathbf{x}; \mathbf{w})\,, \tag{4.7}$$

where $L(\cdot, \cdot)$ is a loss function, which quantifies how well the distribution in Eq. (4.2) is defined by the parameter vector $\mathbf{w}$ matches the labels $\mathbf{y}$.

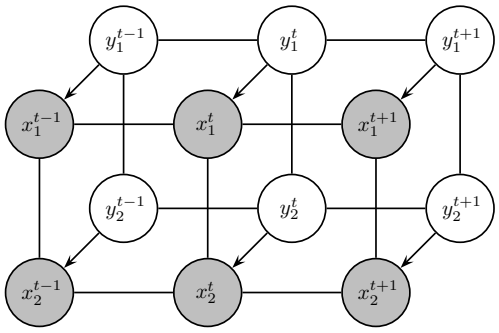


Figure 4.2: Graphical representation of the model. The observed features are represented by $\mathbf{x}$ and the unknown labels are represented by $\mathbf{y}$. Temporal edges exist also between the labels and the observed features across frames.

We select a clique loss function [302], which is defined as the log-likelihood of the posterior probability $p(\mathbf{r}|\mathbf{x};\mathbf{w})$:

$$L(\mathbf{y},\mathbf{x};\mathbf{w}) = -\log p(\mathbf{y}|\mathbf{x};\mathbf{w})\,. \tag{4.8}$$

The loss function is minimized using a gradient-descent optimization method. It can be seen as the empirical risk minimization of the Kullback-Leibler divergence between the true and predicted marginals.

4.2.2 Inference

Having set the parameters $\mathbf{w}$, an exact solution to Eq. (4.1) is generally intractable. For this reason, approximate inference is employed to solve this problem. In this work, LBP [303] is used for computing the marginals using the full graphical model as depicted in Figure 4.2. For comparison purposes and for better insight of the proposed method, we have also tested a variant of the full graphical model by transforming it into a tree-like graph (Figure 4.3). This is accomplished by ignoring the spatial relationship between the observation nodes $\mathbf{x}$ and keeping only the temporal edges between the labels $\mathbf{y}$. In this case, tree-reweighted belief propagation [130] is considered for inference.

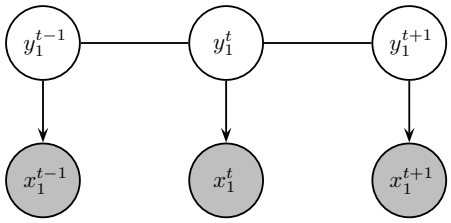


Figure 4.3: Tree-like graphical representation of the model. The observed features are represented by $\mathbf{x}$ and the unknown labels are represented by $\mathbf{y}$.

4.3 Experimental Results

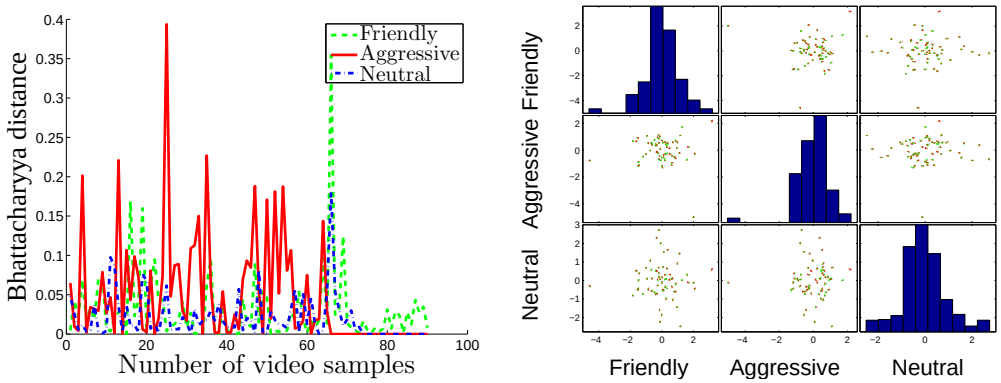
The experiments are applied to the novel *Parliament* dataset [5]. The number of features are kept relatively small in order not to increase the model's complexity. Additionally, to show that the proposed method can perform well, different model variants are compared.

4.3.1 Political Behavior Dataset

To evaluate our method, we collected a set of 228 video sequences, depicting political speeches in the Greek parliament. All behaviors were recorded for 20 different subjects. The videos were acquired with a static camera and contain uncluttered backgrounds. The video sequences were manually labeled with one of three behavioral labels: *friendly* (90 videos), *aggressive* (73 videos), or *neutral* (65 videos). Figure 4.4 depicts some representative frames of the *Parliament* dataset. The subjects express their opinion on a specific law proposal and they adjust their body movements and voice intensity level according to whether they agree with that or not.

Figure 4.4: Sample frames from the proposed *Parliament* dataset. (Top row) Friendly, (middle row) Aggressive, and (bottom row) Neutral.

Each video sequence was manually labeled with one of three behavioral labels according to human perception on kindness and aggressiveness. Figure 4.5 (a) shows the similarity of each class against the other by measuring the Bhattacharyya distance between all pairs of classes. Since the data are multidimensional, viewing slices through lower dimensional subspaces is one way to partially work around the limitation of two or three dimensions. To this end, we employed PCA to project the data onto a three dimensional space and demonstrated how pairs of different classes are distributed in the projected space. Figure 4.5 (b) depicts the distribution of the data of all bivariate scatter plots between all pairs of classes. The plots in the diagonal depict the univariate histogram for each class. Note that all classes are not linearly separable. Within each class, there is a variation in the performance of an action. Each individual exhibits the same



(a) (b)

Figure 4.5: Distribution of classes *friendly*, *aggressive*, and *neutral*. (a) Bhattacharyya distance between classes for all video samples. (b) Distribution of each class against the others (bottom row) after projection onto a common subspace using PCA. The main diagonal shows how data are distributed within each class.

behavior in a different manner by using different body movements. This is an interesting characteristic of the dataset, which makes it challenging.

The videos of the *Parliament* dataset were captured at a resolution of 320×240 pixels at 25 fps and their length is 250 frames. The dataset was annotated by two observers of Greek origin, who watched the videos independently and recorded their labels separately. Disagreement was resolved by a third observer. It is worth noting that the initial two annotators disagreed in only 3% of the videos of the dataset. The observers were asked to categorize the videos with respect to the notions of kindness and aggressiveness according to a general perception of a political speech by a citizen with a Greek mentality as follows. (i) Subjects with large and abrupt body, head and hand movements and high speech signal amplitude are to be labeled as aggressive. This corresponds to statesmen who express strongly their disagreement with the topic discussed or a previous speech given by a political opponent. (ii) Subjects with very small variations in their motion and speech signal amplitude are to be labeled as neutral. This class includes standard political speeches only expressing a point of view without any strong indication (body motion or voice tone) of agreement or disagreement with the topic discussed. (iii) Subjects with large but smooth variations in the pose of their body and hands speaking with a normal speech signal amplitudes are to be labeled as friendly.

We used 5-fold cross validation to split the dataset into training and test sets. Accordingly, the model was learned from 183 videos, while the algorithm was tested on the remaining five videos and the recognition results were averaged over all the examined configurations of training and test sets. Within each class, there is a variation in the performance of an action. Each individual exhibits the same behavior in a different manner by using different body movements. This is an interesting characteristic of the dataset which makes it quite challenging.

Table 4.1: Behavior classification accuracies (%) using the graphical model with only temporal edges (4.2) and the full graphical model (4.1).

Classification Accuracy(%)			
Method	**Friendly**	**Aggressive**	**Neutral**
Tree model (tree-reweighted BP)	100.0	49.2	84.5
Full model (loopy BP)	100.0	60.7	95.8

Table 4.2: Comparison between variants of the proposed method.

Method	**Accuracy(%)**
CRF (unary only)	81.0
CRF (unary no spatio-temporal)	69.7
CRF (pairwise no spatio-temporal)	69.7
Full CRF model	**85.5**

4.3.2 Results and Discussion

We evaluated the proposed model with different variants of the method. First, we compared the full graphical model (see Figure 4.1) with a variant of the method, which considers the graphical model as a tree-like graph (see Figure 4.2). As it can be observed in Table 4.1, the full graphical model performs better than the tree-like graph, which uses only temporal edges between the labels. The second model ignores the spatial relationship between the features and the classification error is increased. Generally, the full graphical model provides strong improvement of more than 8% with respect to the tree model.

In the second set of experiments, we evaluated three variants of the proposed CRF model. First, we used the CRF model with only the unary potentials ignoring the pairwise potentials. The second variant uses only unary potential without the spatio-temporal features. Finally, the third configuration uses the full model without the spatio-temporal pairwise features. The classification results comparing the different models are shown in Table 4.2.

We may observe that the CRF model, which does not usespatio-temporal feature in either the unary potentials or the pairwise potentials, attains the worst performance between the different variants. It is worth mentioning that the first variant, which uses only unary features, performs better than the other two variants, which do not use spatio-temporal features. However, this is not a surprising fact, as in the case of the no spatio-temporal variants the classification is performed for each frame individually ignoring the temporal relationship between consecutive frames. The use of spatio-temporal features appears to lead to better performance than all the other approaches. We also observe that the full CRF model shows significant improvement over all of its variants. The full

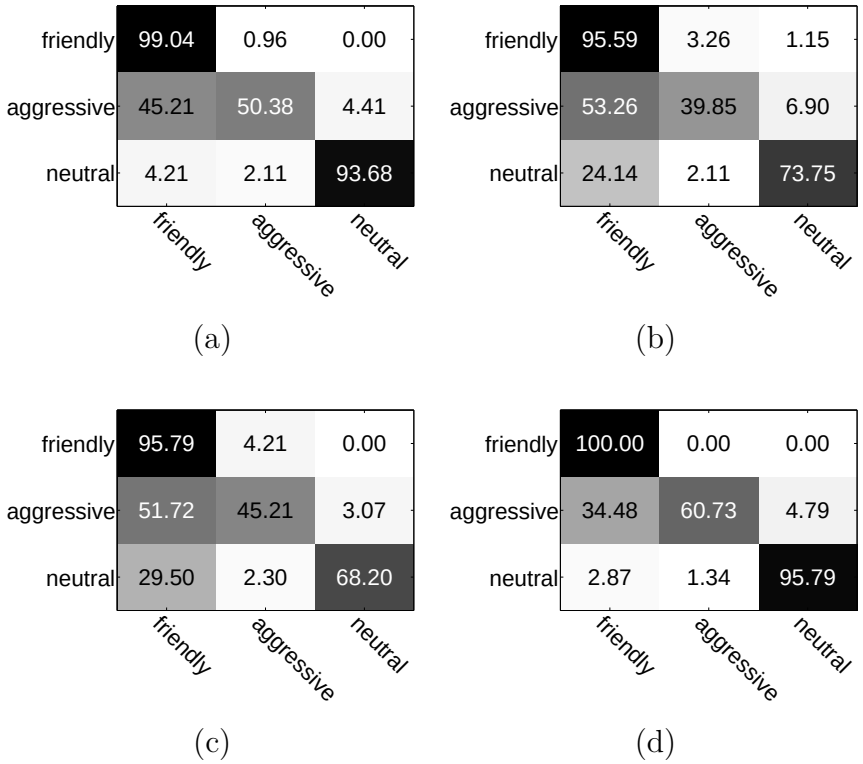


Figure 4.6: Confusion matrices of the classification results for the CRF model employing (a) only unary potentials, (b) only unary potentials without spatio-temporal features, (c) the full model without spatio-temporal pairwise features, and (d) the full model.

CRF model leads also to a significant increase in performance of 85.5%, with respect to the model with no spatio-temporal features. This confirms that temporal and spatial information combined together constitute an important cue for action recognition.

Figure 4.6 illustrates the overall behavior recognition accuracy, where the full CRF model exhibits the best performance in recognizing each of the three behaviors. The main conclusion we can draw from the confusion matrices is that adding temporal edges to the graphical model helps reduce the classification error between the different behavioral states. It is also worth noting that, due to missed and relatively close features in consecutive frames, the classes "friendly" and "aggressive" are often confused as the subject performs similar body movements. Feature selection may be employed to solve this problem.

4.4 Conclusion

In this chapter, a method for recognizing human behaviors in a supervised framework using a CRF model is presented. A new challenging dataset (*Parliament*) was introduced, which captures the behaviors of some politicians in the Greek parliament during their speeches. Several variants of the method were examined reaching an accuracy of 85.5%.

Chapter 5

Identifying Human Behaviors Using Hidden Conditional Random Fields

5.1 Introduction

5.2 Behavior Recognition Using Hidden Conditional Random Fields

5.3 Experimental Results

5.4 Conclusion

5.1 Introduction

Recognizing human behaviors from video sequences is a challenging task [42, 225]. A behavior recognition system may provide information about the personality and psychological state of a person. Its applications vary from video surveillance to human-computer interaction. Human behavior is often expressed as a combination of non-verbal multimodal cues such as gestures, facial expressions and auditory cues. The correlation between cues from different modalities has been shown to improve recognition accuracy [57, 255, 266].

When attempting to recognize human behaviors, one must determine the kinematic states of a person. From psychological point of view, human behaviors may be classified in three types: behavioral, cognitive and social [305]. Our goal is to understand not only social behaviors (e.g., relationships and interactions between people such as hugging or kissing) but also individual behaviors (e.g., expression of personal feelings such as aggressiveness or friendliness).

Factors that can affect human behavior may be decomposed into several components including emotions, moods, actions and interactions with other people. Hence, the recognition of complex actions may be crucial for understanding human behavior. Recognizing human actions that correspond to a specific emotional state of a person or an affective

label such as boredom, or kindness, may help understand social behaviors. The task of learning human behaviors is to identify the psychological state or the social activities of a person taking place in the surroundings [226]. Several affective computing methods [56, 217] used semantic annotations in terms of arousal and valence to capture the underlying affect from multimodal data. However, obtaining affective labels for real world data is a challenging task [212] and it may lead to biased representation of human behaviors.

In this chapter, we address the problem of multimodal data association for human behavior recognition. First, audio and visual data from the video sequences are extracted and then a feature pruning technique is applied to remove redundant features according to the spatiotemporal neighborhood of the features in the video frames. Then, CCA [228] is employed to find the synchronization offset between the audio and video features, such that the correlation between sound emissions and human movements is maximized. Finally, the projected data are concatenated into a new feature vector and are used as input to a chain hidden conditional random field (HCRF) [27] model to capture the interaction across modalities and compute the underlying hidden dynamics between the labels and the features. Our method is also able to cope with videos with varying human poses as feature pruning may reduce the background and discard irrelevant frames. In contrast to most of the multimodal human behavior analysis methods, the combination of feature pruning and early fusion keeps the complexity of our method relatively low, as only one step of classification for estimating human behaviors is required.

The contributions of this work is threefold. First, we developed a supervised multimodal learning framework, for human behavior recognition based on the canonical correlation of audio and visual features. We also proposed a feature selection technique for pruning redundant features, based on the spatio-temporal neighborhood of the visual features that reduced the complexity of the classification algorithm. Finally, we employed an audio-visual synchronization method to temporally align the audio and video features, to better exploit the correlation of the audio-visual features and improve the recognition accuracy.

5.2 Behavior Recognition Using Hidden Conditional Random Fields

We assume that a set of training labels is provided and each video sequence is preprocessed to obtain a bounding box of the human in every frame and each person is associated with a behavioral label. The model is general and can be applied to several behavior recognition datasets. Our method uses HCRFs, which are defined as a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 5.1), as the probabilistic framework for modeling the behavior of a subject in a video. First, audio and visual features are computed in each video frame capturing the roles associated with the bounding boxes. Next, irrelevant visual features are eliminated according to their spatio-temporal relationship of

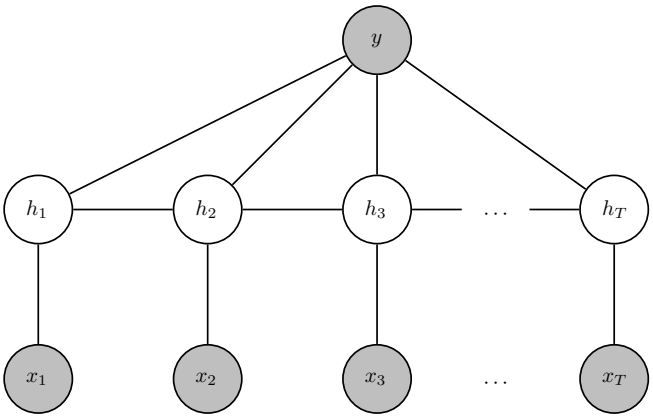


Figure 5.1: Graphical representation of the chain structure model. The grey nodes are the observed features and the unknown labels represented by x and y, respectively. The white nodes are the unobserved hidden variables h.

neighboring features. Then, the synchronization offset between the different modalities is estimated by using CCA. Finally, belief propagation (BP) [169] is applied to estimate the labels.

5.2.1 Multimodal Hidden Conditional Random Fields

We consider a labeled dataset $\mathcal{D} = \{\mathbf{x}_{i,j}, y_i\}_{i=1}^{N}$ with N videos, where $\mathbf{x}_{i,j} = (\mathbf{a}_{i,j}, \mathbf{v}_{i,j})$ is a multimodal observation sequence, which contains audio ($\mathbf{a}_{i,j} \in \mathbb{R}^{n_a \times T}$) and visual data ($\mathbf{v}_{i,j} \in \mathbb{R}^{n_v \times T}$) of length T with $j = 1 \dots T$. For example, $\mathbf{x}_{i,j}$ corresponds to the j^{th} frame of the i^{th} video sequence. Finally, y_i corresponds to a class label defined in a finite label set $\mathcal{Y}$. Our model is applied to all video sequences in the training set. In what follows, we omit indices i and j for simplicity.

It is useful to note that our HCRF model is a member of the exponential family and the probability of the class label given an observation sequence is given by:

$$\begin{aligned} p(y|\mathbf{x};\mathbf{w}) &= \sum_{\mathbf{h}} p(y,\mathbf{h}|\mathbf{x};\mathbf{w}) \\ &= \sum_{\mathbf{h}} \exp\left(E(y,\mathbf{h}|\mathbf{x};\mathbf{w}) - A(\mathbf{w})\right) , \end{aligned} \tag{5.1}$$

where $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters, $\mathbf{h} = \{h_1, h_2, \dots, h_T\}$, with $h_i \in \mathcal{H}$ is a set of latent variables. In particular, the number of latent variables may be different from the number of samples, as h_j may correspond to a substructure in a sample. However, for simplicity we use the same notation. Finally, $E(y,\mathbf{h}|\mathbf{x};\mathbf{w})$ is a vector of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

$$A(\mathbf{w}) = \log \sum_{y'} \sum_{\mathbf{h}} \exp\left(E(y',\mathbf{h}|\mathbf{x};\mathbf{w})\right) . \tag{5.2}$$

Different sufficient statistics $E(y,\mathbf{h}|\mathbf{x};\mathbf{w})$ in (5.1) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configura-

tion of unary and pairwise terms:

$$E(y, \mathbf{h}|\mathbf{x}; \mathbf{w}) = \sum_{j \in \mathcal{V}} \Phi(y, h_j, \mathbf{x}_j; \boldsymbol{\theta}) + \sum_{j,k \in \mathcal{E}} \Psi(y, h_j, h_k; \boldsymbol{\omega}) \,, \tag{5.3}$$

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned and $\Phi(y, h_j, \mathbf{x}_j; \boldsymbol{\theta})$, $\Psi(y, h_j, h_k; \boldsymbol{\omega})$ are the unary and pairwise potentials, respectively.

The unary potential is expressed by:

$$\Phi(y, h_j, \mathbf{x}_j; \boldsymbol{\theta}) = \sum_j \sum_\ell \phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) + \sum_j \phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) \,, \tag{5.4}$$

and it can be considered as a state function, which consists of two different feature functions. The label feature function, which models the relationship between the label y and the hidden variables h_j, is expressed by:

$$\phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) = \sum_{\lambda \in \mathcal{Y}} \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_{1,\ell} \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) \,, \tag{5.5}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if its argument is true and 0 otherwise. The observation feature function, which models the relationship between the hidden variables h_j and the observations $\mathbf{x}_j$, defined by:

$$\phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_2^\top \mathbb{1}(h_j = a) \mathbf{x}_j \,. \tag{5.6}$$

The pairwise potential is a transition function and represents the association between a pair of connected hidden states h_j and h_k and the label y. It is expressed by:

$$\Psi(y, h_j, h_k; \boldsymbol{\omega}) = \sum_{\substack{\lambda \in \mathcal{Y} \\ a,b \in \mathcal{H}}} \sum_\ell \boldsymbol{\omega}_\ell \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) \mathbb{1}(h_k = b) \,. \tag{5.7}$$

5.2.2 Parameter Learning and Inference

Our goal is to assign a test video sequence with a behavioral role by maximizing the posterior probability:

$$y = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}; \mathbf{w}) \,. \tag{5.8}$$

In the training step the optimal parameters $\mathbf{w}^*$ are estimated by maximizing the following loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w}) - \frac{1}{2\sigma^2} \|\mathbf{w}\|^2 \,. \tag{5.9}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}; \mathbf{w})$ and quantifies how well the distribution in (5.1) defined by the parameter vector $\mathbf{w}$ matches the labels y. It can be rewritten as:

$$\log p(y_i|\mathbf{x}_i; \mathbf{w}) = \log \sum_{\mathbf{h}} \exp(E(y, \mathbf{h}|\mathbf{x}_i; \mathbf{w})) - \log \sum_{y', \mathbf{h}} \exp(E(y', \mathbf{h}|\mathbf{x}_i; \mathbf{w})) \,. \tag{5.10}$$

The second term is a Gaussian prior with variance σ^2 and works as a regularizer. The loss function is minimized using a gradient-descent optimization method. More specifically, in our experiments we used the limited-memory BFGS (LBFGS) method [306] to maximize the log-likelihood of the data. Having set the parameters $\mathbf{w}$, the marginal probability is obtained by applying the BP algorithm [25] using the graphical model as depicted in Figure 5.1.

5.2.3 Multimodal Feature Extraction

In this work, we used three different sets of visual features (i.e., STIPs, head orientations, and proxemic features). First, we extract local space-time features at frame rate of 25 fps using a 72-dimensional vector of HoG and 90-dimensional vector of HoF feature descriptors [241] for each STIP [97], which captures the human motion between frames. These features were selected because they can capture salient visual motion patterns in an efficient and compact way.

Feature extraction may be erroneous due to cluttered backgrounds caused by camera motion or changes in illumination and appearance. Reducing the number of irrelevant/redundant features drastically reduces the running time of a learning algorithm and yields a more general concept. For this reason, we adopt a similar technique with Liu *et al.* [4] and we perform feature pruning based on spatial and temporal neighborhood of motion features. The proposed algorithm depends on two factors: (i) the distance between the centers of the feature locations and (ii) the scatter of each feature group in consecutive frames.

Let N_t be the number of features in frame t and N be the total number of features in the video sequence. Let also, μ_t and σ_t^2 be the center and the variance of the feature locations in frame t, respectively. First, we discard those frames where N_t is much larger than the mean number of features in the video sequence. Next, if the ratio of the difference of the means to the standard deviation of feature locations and the number of features between frame t and its neighboring frames $t-1$ and $t+1$ are over a predefined threshold, we select $M_t \leq N_t$ features that lie close to the centers of the feature locations in neighboring frames. A detailed description of the proposed feature pruning algorithm is presented in Algorithm 3. Figure 5.2 depicts some representative examples of the feature pruning technique. Feature pruning may significantly reduce the number of features (Figure 5.5).

In cases where the video sequences are not person-centric, but may contain human interactions (e.g., hugging), STIP features are not adequate. For this reason, we have used head orientation as additional feature. This choice is motivated by the fact that a person who interacts with another is more likely to look at that person than looking at somewhere else. Furthermore, we have also used proxemic features, which capture the spatial and temporal relations between interacting persons detected in the video sequences. This means that interacting persons are in general more probable to lie close to each other (spatially and temporally). Moreover, many audio features have been studied for speaker detection and voice recognition [307]. Mel-frequency cepstral coefficients (MFCCs) [308]

Algorithm 3 Feature pruning

Input: Original features $\mathbf{v}_t$ for frame t.

Output: Pruned features $\mathbf{z}_t$ for frame t.

if $N_t >> \text{mean}(N)$ **then**
    Discard frame t;
end if
if $\left(\frac{\|\mu_{t-1} - \mu_t\|^2}{\sigma_{t-1}^2 + \sigma_t^2} > \varepsilon \ \ \& \ \ \frac{\|\mu_t - \mu_{t+1}\|^2}{\sigma_t^2 + \sigma_{t+1}^2} > \varepsilon \right) \ \ \& \ \ (|N_{t-1} - N_t| > \zeta \ \ \& \ \ |N_t - N_{t+1}| > \zeta)$ **then**
    $j \leftarrow 1$;
    for $i \leftarrow 1$ to N_t **do**
        if $\frac{\|\mu_{t-1} - \mu_t\|^2}{\|\mathbf{v}_{i,t} - \mu_t\|^2} < T \ \ \& \ \ \frac{\|\mu_t - \mu_{t+1}\|^2}{\|\mathbf{v}_{i,t} - \mu_t\|^2} < T$ **then**
            $\mathbf{z}_{j,t} \leftarrow \mathbf{v}_{i,t}$;
            $j \leftarrow j + 1$;
        end if
    end for
end if

Table 5.1: Types of audio and visual features used for human behavior recognition. The numbers in parentheses indicate the dimension of the features.

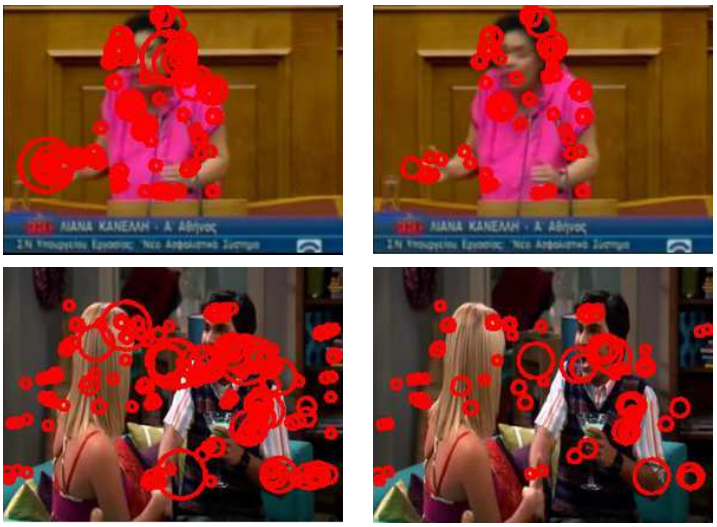
Audio features (39)	**Visual features** (166)
MFCCs (13)	STIP (162)
Delta-MFCCs (13)	Head orientations (2)
Delta-delta-MFCCs (13)	Proxemic (2)

are the most popular and common audio features. We employ the MFCCs features and their first and second order derivatives (delta and delta-delta MFCCs) to form an audio feature vector of dimension 39. Table 5.1 summarizes all audio and visual feature types used in our algorithm.

5.2.4 Audio-Visual Synchronization and Fusion

The purpose of the proposed method is to perform multimodal human behavior recognition by taking into account both visual and audio information. One drawback of combining features of different modalities is the different frame rate that each modality may have. Thus, prior to the fusion step, visual features are interpolated to match the audio frame rate. However, interpolation may harm the synchronization between the audio and visual features, which is necessary to better exploit the correlation between the different modalities. To this end, we propose using CCA to estimate audio-visual synchronization offset and perform the data fusion.

Given a set of zero-mean paired observations $\{(\mathbf{a}_i, \mathbf{v}_i)\}_{i=1}^M$, with $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ and



(a) (b)

Figure 5.2: Representative examples of feature pruning. (a) The original features and (b) the pruned features for the *Parliament* dataset [5] (top row) and the TV human interaction dataset [6] (bottom row). Feature pruning may reduce the number of features by 29% on average.

$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]$, CCA seeks to find two linear transformation vectors $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_v$, such that the correlation $\rho(\boldsymbol{\gamma}_a^\top \mathbf{A}, \boldsymbol{\gamma}_v^\top \mathbf{V})$ between the projections onto these vectors, $\mathbf{a} = \boldsymbol{\gamma}_a^\top \mathbf{A}$ and $\mathbf{v} = \boldsymbol{\gamma}_v^\top \mathbf{V}$ (also known as canonical variates) is maximized:

$$\begin{aligned} \rho(\mathbf{a}, \mathbf{v}) &= \max_{\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_v} \frac{\mathbb{E}[av]}{\sqrt{\mathbb{E}[a]^2 \mathbb{E}[v]^2}} \\ &= \max_{\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_v} \frac{\mathbb{E}[\boldsymbol{\gamma}_a^\top \mathbf{A}\mathbf{V}^\top \boldsymbol{\gamma}_v]}{\sqrt{\mathbb{E}[\boldsymbol{\gamma}_a^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{\gamma}_a] \mathbb{E}[\boldsymbol{\gamma}_v^\top \mathbf{V}\mathbf{V}^\top \boldsymbol{\gamma}_v]}} \\ &= \max_{\boldsymbol{\gamma}_a, \boldsymbol{\gamma}_v} \frac{\boldsymbol{\gamma}_a^\top \Sigma_{av} \boldsymbol{\gamma}_v}{\sqrt{\boldsymbol{\gamma}_a^\top \Sigma_{aa} \boldsymbol{\gamma}_a \boldsymbol{\gamma}_v^\top \Sigma_{vv} \boldsymbol{\gamma}_v}}, \end{aligned} \tag{5.11}$$

where $\mathbb{E}[\cdot]$ is the expected value, $\Sigma_{aa} \in \mathbb{R}^{n_a \times n_a}$ and $\Sigma_{vv} \in \mathbb{R}^{n_v \times n_v}$ are the covariance matrices, respectively, and $\Sigma_{av} \in \mathbb{R}^{n_a \times n_v}$ is the cross-covariance matrix of $\mathbf{A}$ and $\mathbf{V}$.

The solutions for $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_v$ are the eigenvectors corresponding to the largest eigenvalues of $\Sigma_{aa}^{-1}\Sigma_{av}\Sigma_{vv}^{-1}\Sigma_{va}$ and $\Sigma_{vv}^{-1}\Sigma_{va}\Sigma_{aa}^{-1}\Sigma_{av}$, respectively.

The greatest challenge when dealing with audio-visual features is to correctly identify the auditory information that corresponds to the motion of the underlying event. This means, that audio and visual features need to be precisely correlated before data fusion is applied [227, 309]. To this end, we assume that there is a time gap τ, which can be seen as an integer offset of frames between audio and visual streams such that the visual feature vector $\mathbf{v}_t$ in frame t corresponds to the $(t+\tau)^{\text{th}}$ audio feature vector $\mathbf{a}_{t+\tau}$. We assume that the synchronization offset τ may lie in an interval $[-s, s]$. First, we remove

the first and last s frames from the audio signal and compute the audio features in the remaining cropped sequence of length $T - 2s$. Then, we compute the visual features $\mathbf{v}_t$, $t \in [1, 2s + 1]$ in all groups of $T - 2s$ consecutive frames. Finally, CCA is applied between the set of cropped audio features $\mathbf{a}$ and each visual feature group $\mathbf{v}_t$. We select the optimal temporal gap such that the correlation between audio and visual features is maximized according to:

$$\tau = \arg\max_t \lambda_t - (s + 1)\,, \tag{5.12}$$

where λ corresponds to the largest eigenvalue, which is associated with the maximization of the canonical correlation between the audio feature vector and each group of visual features, as the audio feature vector is slid over the visual features. The steps of the audio-visual synchronization algorithm are summarized in Algorithm 4.

Algorithm 4 Audio-visual synchronization

Input: Audio and video streams, time interval $[-s, s]$.

Output: Synchronization offset τ.

1: Delete the first and last s frames from the auditory signal.
2: Compute the audio features in the remaining $T - 2s$ instances of the audio stream.
3: **for all** groups of $T - 2s$ consecutive frames **do**
4: Compute the visual features $\mathbf{v}_t, t \in [1, 2s + 1]$.
5: Estimate the CCA between the cropped audio and the visual features $\mathbf{v}_t$
6: **end for**
7: Estimate the temporal offset τ according to Eq. (5.12).

We now consider the fusion of the audio and visual features $\mathbf{a}$ and $\mathbf{v}$ respectively by projecting these features onto the canonical basis vectors $[\boldsymbol{\gamma}_a^\top, \boldsymbol{\gamma}_v^\top]^\top$ and use this projection for recognition.

5.3 Experimental Results

In what follows, we refer to our *synchronized audio-visual* cues for *activity recognition* method by the acronym SAVAR. The experiments are applied to the novel *Parliament* dataset [5] and the TV human interaction (TVHI) dataset [6]. The number of features is kept relatively small in order not to increase the model's complexity.

5.3.1 Datasets

Parliament [5]: This dataset contains 228 video sequences of political speeches, belonging in three behavioral categories: friendly, aggressive, and neutral. It is described in detail in Chapter 4.

TV human interaction [6]: This dataset consists of 300 video sequences collected from over 20 different TV shows. The video clips contain four kinds of interactions: *hand*

(a) (b) (c)

Figure 5.3: Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral.

shakes, *high fives*, *hugs* and *kisses*, which are equally distributed to the four classes (50 video sequences for each class). Negative examples (e.g., clips that do not contain any of the aforementioned interactions) consist the remaining 100 videos. The length of the video sequences ranges from 30 to 600 frames. The great degree of intra and inter-class diversity between the clips, such as different number of actors in each scene, variations in scale, and changes in camera angle, is an important factor that popularized this dataset for real world evaluation. Some representative frames of the TVHI dataset are illustrated in Figure 5.4.

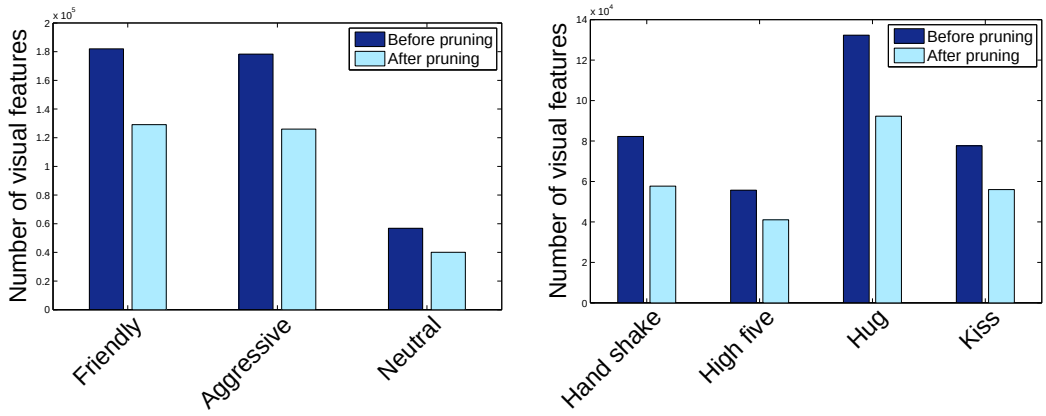
(a) (b) (c) (d)

Figure 5.4: Sample frames from the TVHI dataset. (a) Hand shake, (b) High five, (c) Hug, and (d) Kiss.

In particular, the *Parliament* and the TVHI datasets are representative examples of individual and social behaviors, respectively. The *Parliament* contains examples of behavioral attributes, which may correspond to positive (e.g., friendliness) or negative (e.g., aggressiveness) behaviors. *Passive* is also a possible behavioral state for this dataset. The TVHI dataset on the other hand, models the social behaviors of people in terms of communication/interation with other people. Both kinds of behaviors entail much effort in order to analyze the given information.

5.3.2 Implementation details

We used 5-fold cross validation to split the *Parliament* dataset into training and test sets, and we report the average results over all the examined configurations. Moreover, for the same dataset, we also used the leave-one-speaker-out (LOSO) cross validation, to split training and testing data into two independent sets so that training and testing data



(a) (b)

Figure 5.5: Comparison of the per class number of visual features before and after pruning for (a) the *Parliament* and (b) the TVHI datasets.

may not have utterances from the same speaker. For the evaluation of our method to the TVHI dataset, we used the provided annotations, which are related to the locations of the persons in each video clip including the bounding boxes that contain them, the head orientations of each subject in the clips, the pair of the subjects who interact to each other and the corresponding labels. For comparison purposes, we used the same data split described in [6], which is a 10-fold cross validation. To obtain a bounding box of the human in every frame we used the method described by Dalal and Triggs [85]. Each frame is considered as a grid of overlapping blocks, where HOG features [241] are computed for each block. Finally, a binary SVM classifier is used to identify wether there exists an object or not. The detection window is extracted in all positions and scales and non-maximum suppression is used to detect each object. This method is able to cope with variations in appearance, pose, lighting and complex backgrounds.

The audio signal was sampled at 16 KHz and processed over 10 *ms* using a Hamming window with 25 % overlap. The audio feature vector consisted of a collection of 13 MFCC coefficients along with the first and second derivatives forming a 39 dimensional audio feature vector.

5.3.3 Model Selection

As shown in Figure 5.2, there are many features that are non-informative due to pose variations or complex backgrounds. A comparison of the per class number of visual features before and after pruning using Algorithm 3 for both *Parliament* and TVHI datasets is illustrated in Figure 5.5. It can be observed that the number of visual features before pruning is much higher than the number of visual features after pruning, which indicates that our pruning algorithm may significantly reduce the number of features by 29 % for the *Parliament* dataset and by 27 % for the TVHI dataset on average.

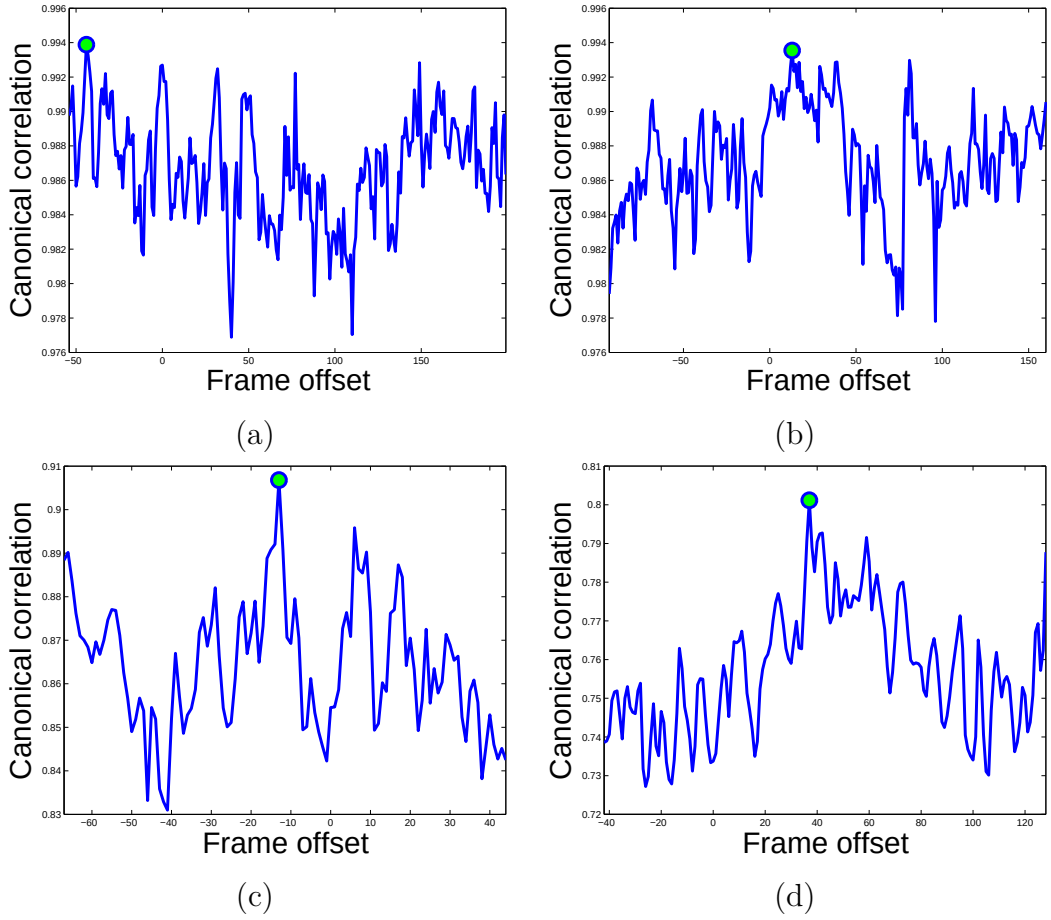
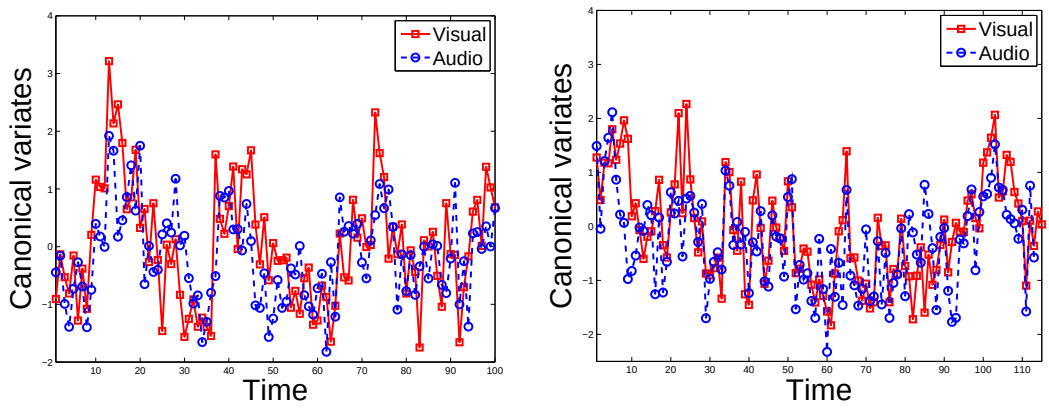


Figure 5.6: Synchronization offsets between audio and video features for some sample video sequences of the *Parliament* (top row) and TVHI (bottom row) datasets. The circle indicates a delay of (a) -44 frames, (b) +13 frames, (c) −13 frames and (d) +37 frames.

To automatically estimate the synchronization offset, such that the correlation between audio and video features is maximized, we used Algorithm 4. Figure 5.6 illustrates the synchronization offset for some randomly selected video sequences by plotting the most significant canonical basis as the visual features slide over the audio features. It is worth noting that, for the synchronization offset, we selected the frame with the maximum correlation. The corresponding canonical bases for the synchronized audio and visual features are depicted in Figure 5.7. The similarity between the audio and visual canonical variates indicates high correlation.

The optimal number of hidden states was automatically estimated based on validation, varying the number of hidden states from three to ten. The L_2 regularization scale term σ was set to $10^k, k \in \{-3, \ldots, 3\}$. Finally, our model was trained with a maximum of 400 iterations for the termination of the LBFGS minimization method.

We compared the SAVAR approach, which uses audio-visual feature synchronization with an HCRF model, SAVAR(A/V sync), with previously reported methods in the litera-



(a) (b)

Figure 5.7: Canonical variates of audio and visual features for two sample videos of (a) the *Parliament* and (b) the TVHI (bottom row) datasets. Notice the high correlation between audio and visual features obtained by the projection.

ture and seven baseline approaches (variants of the proposed method). First, we compared the proposed SAVAR method with an HCRF variant, which does not employ audio-visual feature synchronization prior to the fusion process, SAVAR(A/V no-sync). To show the benefit of audio-visual fusion and synchronization, we compared our SAVAR(A/V sync) method against two HCRF variants, which use only audio, SAVAR(audio), and only visual, SAVAR(visual), features as input, respectively. Moreover, we compared our method with a late fusion technique without using audio-visual synchronization as it is not necessary in late fusion. Information from each modality was learned separately by the HCRF model and then the resulting classification scores were used as input to an SVM model to fuse the results. The parameters of SVM were chosen using cross validation.

A conditional random field model, using four different variants, was also used as a baseline method, to demonstrate the effectiveness of the HCRF model to learn the hidden dynamics between the video clips of different classes. First, synchronized and unsynchronized audio-visual features were used as input to two CRF models comprising two different variants A/V sync CRF and A/V no-sync CRF, respectively. Finally, we trained two CRFs, one with only audio features (audio CRF) and one with only visual (visual CRF) features.

5.3.4 Feature Pruning

The classification accuracy with respect to the number of hidden states before and after feature pruning for both the 5-fold and the LOSO cross validation schemes for the *Parliament* dataset is shown in Table 5.2. It is clear that the model obtained by the proposed algorithm, which uses pruned features, leads to better classification accuracy compared to the model, which uses the un-pruned features for both cross validation schemes. This is due to the fact that the un-pruned visual features may contain outliers and decrease the

recognition accuracy, as the redundant visual features may lead to false estimation of the synchronization offset. Although audio features may improve the overall accuracy of the proposed method, in the case of un-pruned features they do not provide any significant performance as visual features may dominate over the audio features. For LOSO cross validation, and in contrast to the 5-fold scheme, visual features perform better than audio as there exist no utterances from the same speaker, and thus model overfitting, due to existence of redundant information, may be prevented. It is worth mentioning that the accuracy difference between visual and audio cues may be due to the difference in number of features for each modality. The optimal number of hidden states for the 5-fold and LOSO cross validation schemes, which use only audio and only visual data, in the case where feature pruning is used, is six. For the A/V no-sync method the optimal number of hidden states is 10. The number of hidden states remains the same for the LOSO scheme. The optimal number of hidden states for the proposed A/V sync method for the 5-fold scheme is seven, while for the LOSO scheme increases to nine.

Also, Table 5.2 shows the classification results with respect to the number of hidden states when late fusion is applied. It can be seen that the proposed method yields better results than late fusion for both 5-fold and LOSO cross validation schemes. For more than seven hidden states, the results of the proposed method are notably higher than those obtained by late fusion. Although late fusion may work better than the proposed method for a small number of hidden states (3, 5, and 6) for 5-fold cross validation, and 6 hidden states for LOSO cross validation, it is evident that for the majority of number of hidden states the proposed method performs better. Furthermore, even when late fusion outperforms the proposed approach, the improvement is marginal with respect to the improvement obtained by the proposed early fusion approach versus the late fusion for the same number of hidden states. This can be inferred by the fact that the optimal number of hidden states for the proposed 5-fold cross validation scheme is seven and the recognition accuracy is almost 30% higher than corresponding the late fusion approach for the same number of hidden states. Also, for the LOSO cross validation scheme, the recognition accuracy of the proposed method is higher in seven out of eight cases. This might be due to the low number of dimensions that late fusion handles. The proposed method exploits context provided by all modalities and the gain obtained by early fusion corresponds to the synchronized audio-visual cues, as they may be complementary in time. Also, despite the fact that late fusion is a suitable approach for handling multi-modal data, where each modality can be learned separately and differently, we may loose inter-modality dependence, which is crucial for audio-video classification.

The dependence of the classification accuracy and the number of hidden states on the TVHI dataset for both pruned and un-pruned features is shown in Table 5.3. Note that the visual model, which uses the original un-pruned features, performs better than the proposed A/V sync method, which uses pruned visual features, for six and 10 hidden states. This is because the additional visual features may act as outliers and affect the estimation of the true synchronization offset. We can observe that in the case of feature

Table 5.2: Recognition accuracy of the proposed HCRF model with respect to the number of hidden states (h={3 . . . 10}) for the *Parliament* dataset [5] using 5-fold and LOSO cross validation, before feature pruning and after feature pruning.

#Hidden states:	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HCRF before feature pruning using 5-fold cross validation								
A/V sync	29.0	55.7	56.8	64.5	46.3	47.7	51.4	51.0
A/V no-sync	34.6	46.5	55.4	51.0	34.1	44.7	42.0	44.4
Visual	44.9	56.6	47.6	52.9	44.1	40.9	60.8	48.9
HCRF before feature pruning using LOSO cross validation								
A/V sync	67.8	70.0	42.1	52.8	51.8	34.4	35.5	66.5
A/V no-sync	37.1	43.7	47.1	33.4	50.1	44.7	40.9	53.9
Visual	48.4	31.4	47.6	36.4	43.0	43.2	42.6	43.6
HCRF after feature pruning using 5-fold cross validation								
A/V sync	88.1	95.2	85.7	80.2	97.6	95.2	90.5	92.9
A/V no-sync	63.9	66.9	64.4	71.0	69.8	73.8	72.3	78.9
Audio	58.2	71.0	72.7	72.7	54.7	67.1	69.6	67.3
Visual	67.1	57.2	48.2	67.1	15.1	44.9	44.0	59.9
HCRF after feature pruning using LOSO cross validation								
A/V sync	91.0	89.7	94.9	77.1	93.6	94.9	97.4	97.4
A/V no-sync	63.0	59.3	74.9	80.4	76.9	79.2	75.1	89.7
Audio	59.3	63.0	50.0	63.0	51.9	53.7	62.7	50.0
Visual	42.7	63.7	58.2	65.6	60.0	42.7	39.6	58.2
Classification accuracies using late fusion								
Late-fusion (5-fold)	91.1	84.4	89.6	82.9	69.6	72.6	71.9	68.9
Late-fusion (LOSO)	83.3	78.7	83.9	81.5	63.2	67.1	69.3	68.9

pruning the visual model requires seven hidden states to achieve the best classification accuracy. It can also be noted that the audio model achieves the best recognition result by using four hidden states. Although the recognition results for this model are affected by background noise, it is obvious that the combination with the visual information can significantly improve the recognition rate. The A/V no-sync method requires eight hidden states, while the proposed A/V sync method uses nine hidden states to reach the best recognition accuracy. The number of hidden states depends not only on the number of the classes in a specific dataset, but also on the variety of the features used.

Table 5.3 demonstrates also the classification results, when late fusion is applied.

Table 5.3: Recognition accuracy of the proposed HCRF model with respect to the number of hidden states (h=$\{4 \dots 10\}$) the TVHI dataset [6] before feature pruning and after feature pruning.

#Hidden states:	**4**	**5**	**6**	**7**	**8**	**9**	**10**
HCRF before feature pruning							
A/V sync	40.6	60.9	46.9	43.8	53.1	54.7	54.7
A/V no-sync	39.1	42.2	40.6	32.8	46.9	51.6	35.9
Visual	35.9	37.5	48.4	42.2	29.9	35.9	60.9
HCRF after feature pruning							
A/V sync	53.1	79.7	70.3	73.4	73.4	81.3	76.6
A/V no-sync	46.9	53.1	35.9	56.6	60.9	54.7	42.2
Audio	35.9	34.4	29.7	28.1	28.1	32.8	23.4
Visual	28.1	50.0	59.4	60.9	37.5	35.9	57.8
Classification accuracies using late fusion							
Late-fusion	80.1	75.0	73.4	75.0	71.8	78.1	76.5

Although in three out of seven cases, the late fusion scheme was able to improve the classification results, the proposed early fusion method performed better for the majority of the different number of hidden states. This is due to the heterogeneity of the different modalities and the confidence scores of each classifier, which may affect the discriminative ability of the SVM classifier as it may assign larger weights to scores that are less prominent.

Taking a closer look at the visual model, we can see that the number of hidden states plays a crucial role in the recognition process; when the hidden states are increased from six to seven, recognition accuracy falls drastically from 67.1 % to 15.1 % for the *Parliament* dataset and from 60.9 % to 37.5 % for the TVHI dataset. In order to estimate the optimal number of hidden states we used cross validation. The reason for reporting the classification accuracies for all hidden states and not only for the optimal configuration is to demonstrate the behavior of the method with respect to the different number of hidden states and the cross validation schemes. It is also worth noting that 5-fold and LOSO cross validation schemes do not achieve the best accuracy for the same number of hidden states, which leads us to the conclusion that knowing in advance the optimal number of hidden states is not an easy task. Moreover, for both datasets, the optimal number of hidden states for each method with respect to the recognition accuracy is depicted in bold in Tables 5.2 and 5.3. When the same accuracy is achieved for more than one hidden states, the smallest number is considered to be the optimal. However, a larger

Table 5.4: Classification results on the *Parliament* dataset [5].

Method	**Accuracy (%)**			
	Audio	**Visual**	**A/V no-sync**	**A/V sync**
Vrigkas *et al.* [5]	N/A	**85.5**	N/A	N/A
CRF [26]	50.3	78.1	67.6	83.7
SAVAR-5-fold	**72.7**	67.1	78.9	**97.6**
SAVAR-LOSO	62.2	65.5	**89.7**	97.4

number of hidden states may lead to a severe overfitting of the model. In this case, the regularization term in Eq. (5.9) may act as a preventer however, tuning the regularization parameters may be difficult and thus, overfitting may not be perfectly eliminated. It is also worth mentioning that both the *Parliament* and the TVHI datasets hold strong intra-class variabilities as certain classes are often confused because the subject performs similar body movements. This confirms that audio and visual information combined together constitute an important cue for action recognition.

5.3.5 Comparison of Learning Frameworks

Tables 5.4 and 5.5 report the classification accuracy on the *Parliament* dataset, for both 5-fold and LOSO cross validation schemes, and the TVHI datasets, respectively. We compare our SAVAR(A/V sync) method with the seven baseline methods and include previous results for each dataset reported in the literature. The results indicate that our approach captures the hidden dynamics between the clips (i.e., the interaction between an arm lift and the raise in the voice). It is clear that HCRFs outperform CRFs when multimodal data are used for the recognition task. Notably, our approach achieves very high recognition accuracy for the *Parliament* dataset (97.6 %), when 5-fold cross validation is used. Comparable results are also provided by the LOSO cross validation scheme as the recognition accuracy is only by 0.2 % lower than the 5-fold cross validation counterpart method. Note that for the SAVAR(A/V no-sync) variant, when LOSO scheme is used, the classification accuracy is by approximately 12 % higher than the corresponding 5-fold cross validation method. Also, when the 5-fold cross validation scheme is employed, SAVAR(audio) performs better than SAVAR(visual) as training data may have utterances from the same speaker. For the LOSO scheme, where the same speaker is excluded from the training data, visual features perform by approximately 3 % better than the acoustic.

The method in [5] employs a fully connected CRF model, where not only the labels but also the observation samples are associated to each other between consecutive frames. That is, the method in [5] assigns a distinct label to each frame, which makes it more suitable to cope with un-segmented videos (i.e., videos with more than one class labels). On the other hand, this property significantly increases the complexity of the method, which makes it quite difficult to use for large video clips.

Also, Table 5.5 demonstrates that the SAVAR approach performs significantly higher

Table 5.5: Classification results on the TVHI dataset [6].

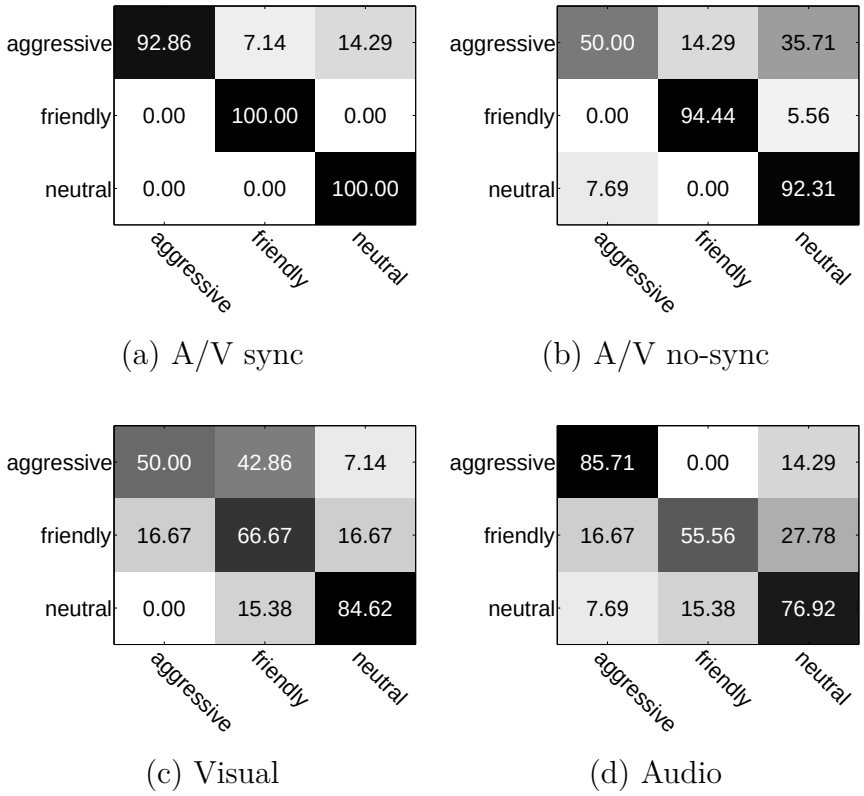
Method	**Accuracy (%)**			
	Audio	**Visual**	**A/V no-sync**	**A/V sync**
Patron-Perez *et al.* [6]	N/A	54.7	N/A	N/A
Li *et al.* [103]	N/A	**68.0**	N/A	N/A
Yu *et al.* [99]	N/A	66.2	N/A	N/A
Gaidon *et al.* [310]	N/A	55.6	N/A	N/A
Marín-Jiménez *et al.* [58]	**48.5**	46.0	54.5	N/A
CRF [26]	36.7	38.7	49.5	52.8
SAVAR	35.9	60.9	**60.9**	**81.3**

than other methods proposed in the literature for the TVHI dataset, by achieving an accuracy of 81.3 %, which is remarkably higher than the best recognition accuracy (68 %) for this dataset achieved by Li *et al.* [103], when only visual features are used, and the best recognition accuracy (54.5 %) achieved by Marín-Jiménez *et al.* [58], when audio and visual features are combined together. It is also worth noting that the SAVAR(visual) and the SAVAR(A/V no-sync) models achieve the same recognition accuracy for this dataset, indicating how important the audio-visual synchronization is for the recognition task, as the unsynchronized multimodal data may not provide any further information to the overall process. For the methods [6, 58, 99, 103, 310] the standard deviations of the classification accuracies are not provided in the original papers and thus, they are not included in Table 5.5.

The resulting confusion matrices of the proposed method for the optimal number of hidden states for the *Parliament* dataset using 5-fold and LOSO cross validation, are depicted in Figures 5.8 and 5.9. The proposed SAVAR(A/V sync) method has significantly small classification errors between different classes, when is compared to the other variants, for both 5-fold and LOSO cross validation schemes. The SAVAR(A/V no-sync) variant has also good classification results and particularly, for the LOSO cross validation scheme, it can perfectly recognize the classes *friendly* and *neutral*. It is also interesting to observe that the different classes for the SAVAR(visual) and the SAVAR(audio) variants may be strongly confused, which emphasizes the fact that when combining audio and visual information together we are able to better separate the emotional states of a person.

Finally, the confusion matrices for the TVHI dataset are shown in Figure 5.10. The smallest classification error between classes belongs to the proposed SAVAR(A/V sync) method. Note that the different classes may be strongly confused as the TVHI dataset has large intra-class variability. Especially, the SAVAR(audio) variant has the largest classification error among all other variants as all classes are confused with the class *kiss*. This is due to the fact that in class *kiss* the audio information may serve as outlier since it contains background sounds.

In order to provide a statistical evidence of the recognition accuracy, we computed the p-values of the obtained results with respect to the compared methods. The null



(a) A/V sync (b) A/V no-sync

(c) Visual (d) Audio

Figure 5.8: Confusion matrices for the classification results of the proposed SAVAR approach for the *Parliament* dataset [5], after feature pruning, using 5-fold cross validation.

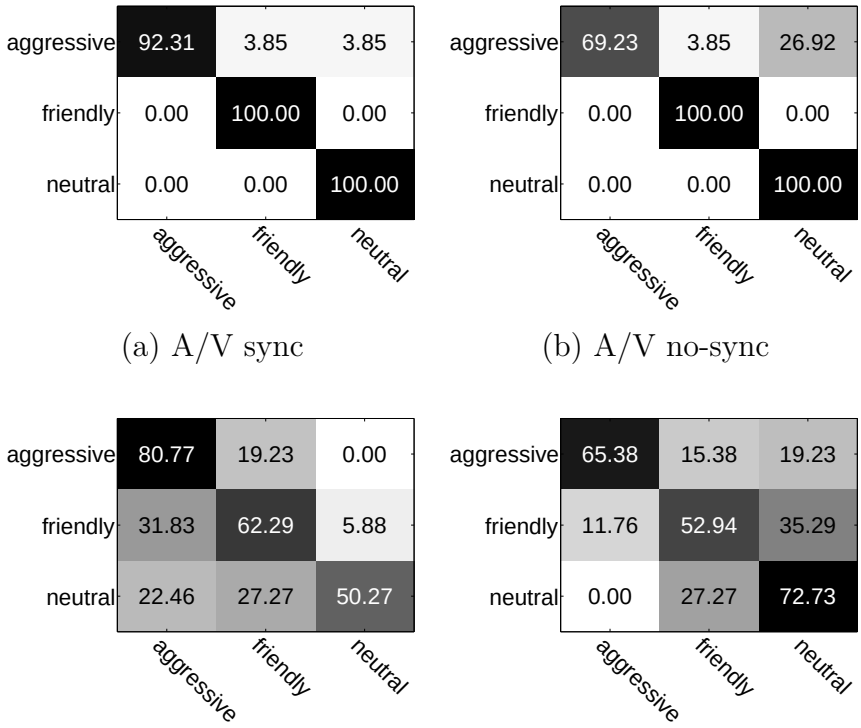
Table 5.6: p-values of the proposed method for the *Parliament* dataset [5].

Method	**SAVAR-5-fold**	**SAVAR-LOSO**
Vrigkas *et al.* [5]	0.0200	0.0058
CRF [26]	0.0137	0.0047

hypothesis was defined as: the mean performances of the proposed model are the same as those of the state-of-the-art methods; and the alternative hypothesis was defined as: the mean performances of the proposed model are higher than those of the state-of-the-art methods. For the assessment of the statistical significance, we used paired t-tests with statistical significance threshold $p < 0.05$ for all experiments.

For the *Parliament* dataset (Table 6), we may observe that the SAVAR-5-fold and SAVAR-LOSO approaches reject the null hypothesis as all values are greater than the critical value (95% of significance level). For the TVHI dataset (Table 7) the null hypothesis is rejected for the majority of the cases. That is, for four out of six cases the p-values were less than the significance level of 0.05. Therefore, we may conclude that the null hypothesis can be rejected and the improvements obtained by our model are statistically significant.

The main strength of the proposed method is that it achieves remarkably good classifi-



(a) A/V sync (b) A/V no-sync

(c) Visual (d) Audio

Figure 5.9: Confusion matrices for the classification results of the proposed SAVAR approach for the *Parliament* dataset [5], after feature pruning, using LOSO cross validation.

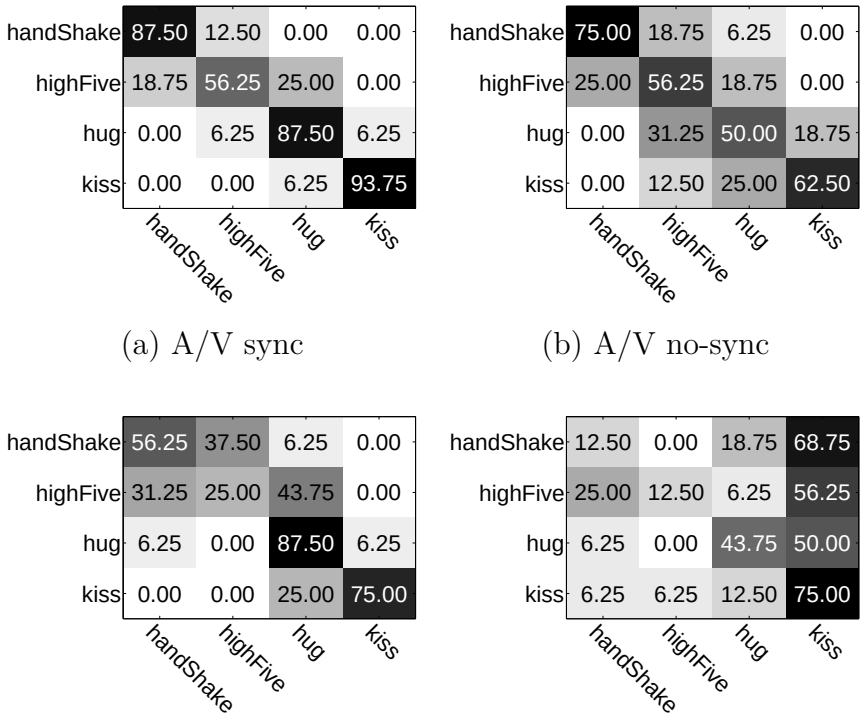
Table 5.7: p-values of the proposed method for the TVHI dataset [6].

Method	**SAVAR**
Patron-Perez *et al.* [6]	0.0012
Li *et al.* [103]	0.1239
Yu *et al.* [99]	0.0620
Gaidon *et al.* [310]	0.0015
Marín-Jiménez *et al.* [58]	0.0002
CRF [26]	0.0007

cation results when synchronized multimodal features are used compared with the results reported in the literature for the same datasets. Additionally, it keeps the number of visual features relatively small by pruning irrelevant features, thus reducing the computational burden of the method.

5.4 Conclusion

In this chapter, the problem of human behavior recognition in a supervised framework using a HCRF model with multimodal data was studied. Specifically, audio features were



(a) A/V sync (b) A/V no-sync

(c) Visual (d) Audio

Figure 5.10: Confusion matrices for the classification results of the proposed SAVAR approach for the TVHI dataset [6], after feature pruning.

jointly used with the visual information to take into account natural human actions. To prune redundant features, a feature selection technique, based on the spatio-temporal neighborhood of each feature in a video clip, was proposed. This has helped reduce the number of features and sped up the learning process.

Furthermore, a method for multimodal feature synchronization and fusion using CCA was also proposed. The evaluation of the proposed method, showed that a moving subject is highly correlated with the auditory information, as human behaviors are characterized by complex actions of movements and sound emissions. The experimental results indicated that the exact synchronization of multimodal data before feature fusion ameliorates the recognition performance. In addition, the combination of audio and visual cues may lead to better understanding of human behaviors. The main strength of this method is that the proposed multimodal fusion approach is general and it can be applied to several types of features for recognizing realistic human actions.

According to the obtained results, the proposed SAVAR method, when it is used with synchronized audio-visual cues, achieves notably higher performance than all the compared classification schemes. This could be seen as an additional characteristic of our model to discriminate between similar classes, when multimodal data is used. Nonetheless, when only one modality was used, the method seemed to have difficulties in efficiently recognizing human behaviors, but it could yield comparable results to the multimodal

SAVAR method. That is, although the combination of audio and visual cues could constitute a strong attribute for discriminating between different classes, each modality separately was unable to capture the variation in temporal patterns of the input data. The proposed method was also able to deal with natural video sequences. The visual feature pruning process could significantly reduce the amount of irrelevant features extracted in each frame, and considerably increased the classification performance with respect to all methods that do not incorporate feature pruning.

Chapter 6

Human Activity Recognition Using Robust Adaptive Privileged Probabilistic Learning

6.1 Introduction

6.2 Robust Privileged Probabilistic Learning

6.3 Experimental Results

6.4 Conclusion

6.1 Introduction

Recent advances in computer vision such as video surveillance and human-machine interactions [311, 312] rely on machine learning techniques trained on large scale human annotated datasets. However, training data may not always be available during testing and learning using privileged information (LUPI) [313] has been used to overcome this problem. The insight of privileged information is that one may have access to additional information about the training samples, which is not available during testing.

Consequently, classification models may often suffer from "structure imbalance" between training and testing data, which may be represented by the LUPI paradigm. Since the additional features are considered more informative than the initial features, the lack of such information during testing is interpreted as an imbalance between training and testing data. This learning technique simulates a real-life learning condition, when a student learns from his/her teacher, where the latter provides the student with additional knowledge, comments, explanations, or rewards in class. Subsequently, the student should be able to face any problem related to what he/she has learned without the help

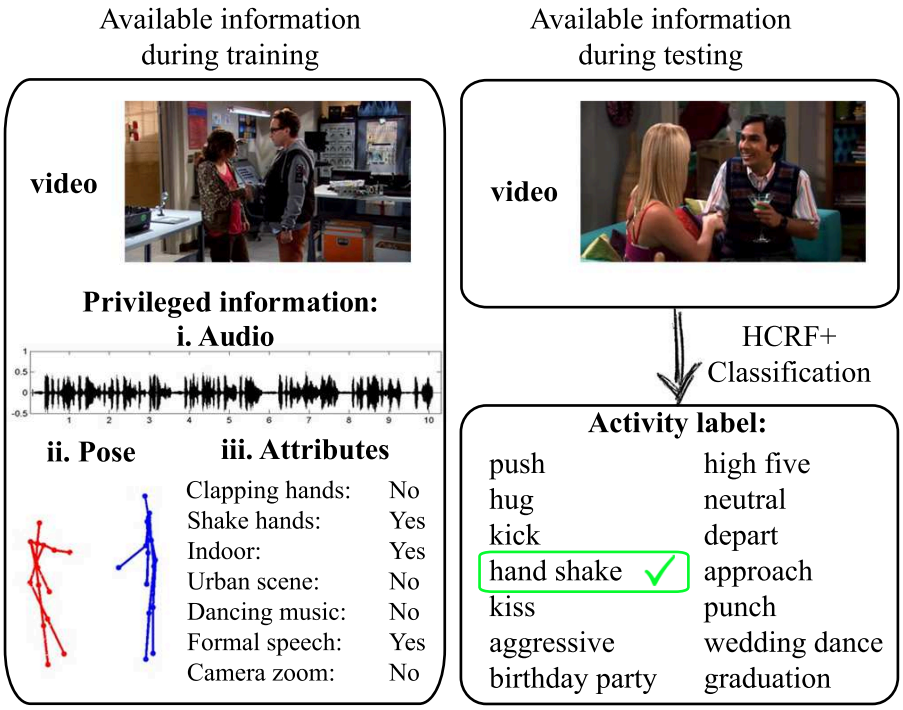


Figure 6.1: Robust learning using privileged information. Given a set of training examples and a set of additional information about the training samples (left) our system can successfully recognize the class label of the underlying activity without having access to the additional information during testing (right). We explore three different forms of privileged information (e.g., audio signals, human poses, and attributes) by modeling them with a Student's t-distribution and incorporating them into the HCRF+ model.

of the teacher. Taking advantage of this learning model, the LUPI framework has also been used in several machine learning applications such as boosting [314], clustering [315], facial expression recognition [316] and textual description [317].

The problem of human activity understanding using privileged knowledge is on its own a very challenging task. Since privileged information is only available during training, one should combine both original and privileged information into a unified classifier to predict the true class label. That is, the learning model should be able to combine both types of information to enhance the classification accuracy by learning a better estimate of model parameters. However, it is quite difficult to identify the most useful information to be used as privileged as the lack of informative data or the presence of misleading information may influence the performance of the model by introducing bias.

We address these issues by presenting a new probabilistic approach, which is able to learn human activities by exploiting additional information about the input data, that may reflect on natural or auxiliary properties about classes and members of the classes of the training data (see Fig. 6.1) and it is used for training purposes only and not for predicting the true classes (where, in general, this information is missing). It is worth noting that the proposed methodology is not limited to the use of a specific form of privileged information, but it is general and may handle any form of additional data. We

also discuss how the privileged information can be used for recognizing human activities when the input may consist of data from different modalities.

Within this framework, we employ a new learning method based on hidden conditional random fields (HCRFs) [27], called HCRF+, which is able to capture the underlying hidden dynamics between the labels and the features in a way that is independent of the learning function involving the additional feature set. In particular, the proposed HCRF+ method differentiates from previous approaches, which may also use the LUPI paradigm, by incorporating privileged information in a supervised probabilistic manner, which facilitates the training process by learning the conditional probability distribution between human activities and observations. We show that both maximum likelihood and maximum margin learning methods may be used to estimate the model's parameters. Furthermore, we introduce a novel technique for automatic estimation of the optimal regularization parameters for the learning process for both maximum likelihood and max-margin approaches. The method is adaptive as the regularization parameters are computed from the training data through a self-training procedure.

Moreover, our method can efficiently manage dissimilarities in input data, which may correspond to noise, missing data, or outliers, using a Student's t-distribution to model the conditional probability of the privileged information. Such dissimilarities may harm the classification accuracy and lead to excessive sensitivity when input data is small or contains large intra-class variations. In particular, the use of Student's t-distribution is justified by the property that it has heavier tails than a standard Gaussian distribution, thus providing robustness to outliers [318].

The main contributions of this work can be summarized in the following points: (i) a human activity recognition method is proposed, which exploits privileged information in a probabilistic manner by introducing a novel classification scheme based on hidden conditional random fields to deal with missing or incomplete data during testing; (ii) both maximum likelihood and maximum margin approaches are incorporated into the proposed HCRF+ model; (iii) a novel technique for adaptive estimation of the regularization term during the learning process is introduced by incorporating both privileged and original data. (iv) contrary to previous methods, which may be sensitive to outlying data measurements, a robust framework for recognizing human activities is intergraded by employing a Student's t-distribution to attain robustness against outliers; (v) the generic nature of our approach is emphasized with the use of samples from different modalities (e.g., data samples may contain information from audio and visual cues) as no further assumption about the kind of training information is made.

6.2 Robust Privileged Probabilistic Learning

We assume that a set of training labels is provided and each video sequence is preprocessed to obtain a bounding box of the human in every frame and each person is associated with a behavioral label. The model is general and can be applied to several

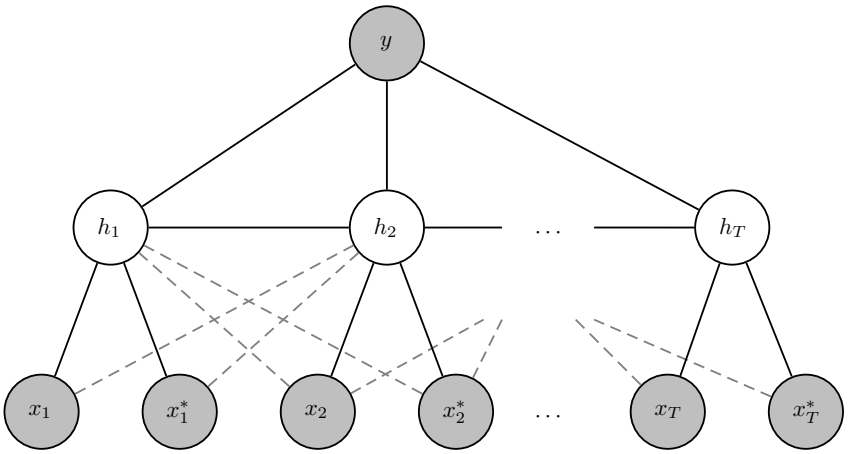


Figure 6.2: Graphical representation of the chain structure model. The grey nodes are the observed features (x_i), the privileged information (x_i^*), and the unknown labels (y), respectively. The white nodes are the unobserved hidden variables (h).

activity recognition datasets. Our method uses HCRFs, which are defined by a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 6.2), as the probabilistic framework for modeling the behavior of a subject in a video.

During training, a classifier and the mapping from observations to the label set for the different configurations are learned. In testing, a probe sequence is classified into its respective state using loopy belief propagation (LBP) [303].

6.2.1 HCRF+ Model Formulation

We consider a labeled dataset with N video sequences, which instead of paired input-output samples $\mathcal{D} = \{(\mathbf{x}_{i,j}, y_i)\}_{i=1}^N$ it consists of triplets $\mathcal{D} = \{(\mathbf{x}_{i,j}, \mathbf{x}_{i,j}^*, y_i)\}_{i=1}^N$, where $\mathbf{x}_{i,j} \in \mathbb{R}^{M_\mathbf{x} \times T}$ is an observation sequence of length T with $j = 1 \dots T$. For example, $\mathbf{x}_{i,j}$ might correspond to the j^{th} frame of the i^{th} video sequence. Furthermore, y_i corresponds to a class label defined in a finite label set $\mathcal{Y}$. In the context of robust learning using a privileged information paradigm, additional information about the observations $\mathbf{x}_i$ is encoded in a feature vector $\mathbf{x}_{i,j}^* \in \mathbb{R}^{M_{\mathbf{x}^*} \times T}$. Such privileged information is provided only at the training step and it is not available during testing. Note that we do not make any assumption about the form of the privileged data.

In particular, $\mathbf{x}_{i,j}^*$ does not necessarily share the same characteristics with the original data, but is rather computed as a very different kind of information, which may contain verbal and/or non-verbal multimodal cues such as (i) visual features, (ii) semantic attributes, (iii) textual descriptions of the observations, (iv) image/video tags, (v) human poses, and (vi) audio cues. The goal of LUPI is to use the privileged information $\mathbf{x}_{i,j}^*$ as a medium to construct a better classifier for solving practical problems than one would learn without it. In what follows, we omit indices i and j for simplicity.

The HCRF+ model is a member of the exponential family and the probability of the

class label given an observation sequence is given by:

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) &= \sum_{\mathbf{h}} p(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) \\ &= \sum_{\mathbf{h}} \exp\left(E(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) - A(\mathbf{w})\right) , \end{aligned} \tag{6.1}$$

where $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters, and $\mathbf{h} = \{h_1, h_2, \ldots, h_T\}$, with $h_j \in \mathcal{H}$ being a set of latent variables. In particular, the number of latent variables may be different from the number of samples, as h_j may correspond to a substructure in an observation. Moreover, the features follow the structure of the graph, in which no feature may depend on more than two hidden states h_j and h_k [27]. This property not only captures the synchronization points between the different sets of information of the same state, but also models the compatibility between pairs of consecutive states. We assume that our model follows the first-order Markov chain structure (i.e., the current state affects the next state). Finally, $E(y, \mathbf{h}|\mathbf{x}; \mathbf{w})$ is a vector of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

$$A(\mathbf{w}) = \log \sum_{y'} \sum_{\mathbf{h}} \exp\left(E(y', \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w})\right) . \tag{6.2}$$

Different sufficient statistics $E(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ in (6.1) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$E(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) = \sum_{j \in \mathcal{V}} \Phi(y, h_j, \mathbf{x}_j, \mathbf{x}_j^*; \boldsymbol{\theta}) + \sum_{j,k \in \mathcal{E}} \Psi(y, h_j, h_k; \boldsymbol{\omega}) , \tag{6.3}$$

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned. Moreover, the potential functions correspond to the structure of the graphical model as illustrated in Fig. 6.2. For example, a unary potential does not depend on more than two hidden variables h_j and h_k, and a pairwise potential may depend on h_j and h_k, which means that there must be an edge (j, k) in the graphical model.

The unary potential is expressed by:

$$\Phi(y, h_j, \mathbf{x}_j, \mathbf{x}_j^*; \boldsymbol{\theta}) = \sum_j \sum_\ell \phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) + \sum_j \phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) + \sum_j \phi_3(h_j, \mathbf{x}_j^*; \boldsymbol{\theta}_3) , \tag{6.4}$$

and it can be seen as a state function, which consists of three different feature functions. The label feature function, which models the relationship between the label y and the hidden variables h_j, is expressed by:

$$\phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) = \sum_{\lambda \in \mathcal{Y}} \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_{1,\ell} \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) , \tag{6.5}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if its argument is true and 0 otherwise. The number of the label feature functions is $|\mathcal{Y}| \times |\mathcal{H}|$. The observation

feature function, which models the relationship between the hidden variables h_j and the observations $\mathbf{x}_j$, is defined by:

$$\phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_2^\top \mathbb{1}(h_j = a)\mathbf{x}_j \,. \tag{6.6}$$

The number of the observation feature functions is considered to be $|\mathcal{Y}| \times |M_{\mathbf{x}}|$. Finally, the privileged feature function, which models the relationship between the hidden variables h_j and the privileged information $\mathbf{x}_j^*$, has $|\mathcal{Y}| \times |M_{\mathbf{x}^*}|$ number of functions and is defined by:

$$\phi_3(h_j, \mathbf{x}_j^*; \boldsymbol{\theta}_3) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_3^\top \mathbb{1}(h_j = a)\mathbf{x}_j^* \,. \tag{6.7}$$

The pairwise potential is a transition function and represents the association between a pair of connected hidden states h_j and h_k and the label y. It is expressed by:

$$\Psi(y, h_j, h_k; \boldsymbol{\omega}) = \sum_{\substack{\lambda \in \mathcal{Y} \\ a,b \in \mathcal{H}}} \sum_{\ell} \boldsymbol{\omega}_\ell \mathbb{1}(y = \lambda)\mathbb{1}(h_j = a)\mathbb{1}(h_k = b) \,. \tag{6.8}$$

The number of the transition functions is $|\mathcal{Y}| \times |\mathcal{H}|^2$. Note that the HCRF+ model keeps a transition matrix for each label.

6.2.2 Maximum Likelihood Learning

In the training step the optimal parameters $\mathbf{w}^*$ are estimated by maximizing the following loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \frac{1}{\lambda_i} \log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \frac{1}{2\sigma^2} \|\mathbf{w}\|^2 \,. \tag{6.9}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ and quantifies how well the distribution in (6.1) defined by the parameter vector $\mathbf{w}$ matches the labels y, while λ is a tuning parameter. It can be rewritten as:

$$\log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) = \log \sum_{\mathbf{h}} \exp(E(y, \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})) - \log \sum_{y' \neq y, \mathbf{h}} \exp(E(y', \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})) \,. \tag{6.10}$$

The second term in Eq. (6.9) is a Gaussian prior with variance σ^2 and works as a regularizer. The use of hidden variables makes the optimization of the loss function non-convex, thus, a global solution is not guaranteed and we can estimate $\mathbf{w}^*$ that are locally optimal. The loss function is optimized using a gradient-descent optimization method. More specifically, in our experiments we used the limited-memory BFGS (LBFGS) method [306] to minimize the negative log-likelihood of the data.

6.2.3 Maximum Margin Learning

We can easily alter the optimization problem of the loss function defined in Eq. (6.9) into a max-margin problem by substituting the summation over the hidden states and the

labels in Eq. (6.10) with maximization [69]. The goal is to maximize the margin between the score of the correct label and the score of the other labels. To learn the parameters $\mathbf{w}^*$ we need to minimize a loss function of the form:

$$\begin{aligned} L(\mathbf{w}) &= \sum_{i=1}^{N} \frac{1}{\lambda_i} \xi_i + \frac{1}{2\sigma^2} \|\mathbf{w}\|^2 \\ \text{s.t. } \max_{\mathbf{h}} E(y, \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) &- \max_{y' \neq y, \mathbf{h}} E(y', \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) \leq \xi_i - 1 \text{ and } \xi_i > 0, \ \forall i \,. \end{aligned} \tag{6.11}$$

Parameter λ is the trade-off between the classification accuracy and the regularization term. Note that although we add slack variables ξ to max-margin optimization, they eventually vanish. We do not estimate the slacks, but we replace them with the Hinge loss error [319] that penalizes the loss when the constraints in Eq. (6.11) are violated:

$$\ell_i(\mathbf{w}) = \max(0, 1 + (\max_{\mathbf{h}} E(y, \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \max_{y' \neq y, \mathbf{h}} E(y', \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}))) \,. \tag{6.12}$$

The optimization problem in (6.11) is equivalent to the optimization of the following unconstrained problem:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \frac{1}{\lambda_i} \ell_i(\mathbf{w}) + \frac{1}{2\sigma^2} \|\mathbf{w}\|^2 \,. \tag{6.13}$$

However, the quantity $\max(0, \cdot)$ is not differentiable and thus, Eq. (6.11) it is hard to solve. To overcome this problem we adapt the bundle method [320], which uses sub-gradient descent optimization algorithm.

6.2.4 Estimation of Regularization Parameters

Both maximum likelihood and max-margin loss functions introduce regularization parameters that control data fidelity and these regularization parameters in Eqs. (6.9) and (6.13), may be obtained in closed form. Here, we examine the case of maximum likelihood optimization as the estimation of the regularization parameters for the max-margin optimization is equivalent. We can rewrite the loss function in Eq. (6.9) as the sum of individual smoothing functionals for each of the training samples N:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \left\{ \log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 \right\} , \tag{6.14}$$

where $\alpha_i(\mathbf{w}) \equiv \dfrac{\lambda_i}{2\sigma^2}$.

In general, the choice of the regularization parameter for the optimization of the loss function should be a function of model parameters $\mathbf{w}$. We consider a linear function $f(\cdot)$ between α_i and each term of the loss function:

$$\begin{aligned} \alpha_i(\mathbf{w}) &= f\Big(\sum_{i=1}^{N} \left\{ \log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 \right\} \Big) \\ &= \frac{1}{\gamma_i} \Big\{ \sum_{i=1}^{N} \left\{ \log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 \right\} \Big\} , \end{aligned} \tag{6.15}$$

where γ_i is determined by the sufficient conditions for convergence:

$$\frac{1}{\gamma_i} < \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w})\|\mathbf{w}\|^2 \,. \tag{6.16}$$

We assume that the privileged information is more informative for classifying human actions than the regular information. Note that, this is the intuition of using of privileged information as additional features for classification purposes and it may hold for most of the cases. Thus, the loss of classifying human actions directly from $\mathbf{x}$ should be lower than classifying from both $\mathbf{x}$ and $\mathbf{x}^*$:

$$\log p(y_i|\mathbf{x}_i; \mathbf{w}) < \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) \,. \tag{6.17}$$

We can then relax the problem and consider that Eq. (6.16) is satisfied when:

$$\frac{1}{\gamma_i} = \log p(y_i|\mathbf{x}_i; \mathbf{w}) \,. \tag{6.18}$$

Thus, the regularization parameter α_i for the loss function is given by:

$$\begin{aligned} \alpha_i(\mathbf{w}) &= \frac{\log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})}{\dfrac{1}{\gamma_i} + \|\mathbf{w}\|^2} \\ &= \frac{\log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})}{\log p(y_i|\mathbf{x}_i; \mathbf{w}) + \|\mathbf{w}\|^2} \,. \end{aligned} \tag{6.19}$$

The regularization parameter α_i may act as as the within-classification balance between data and model parameters. In each step of the optimization process we adaptively update the regularization parameter α_i providing robustness to the trade-off between the regularization terms.

Similarly, the regularization parameter α_i for the loss function in the case of max-margin optimization is given by:

$$\alpha_i(\mathbf{w}) = \frac{\ell_i(\mathbf{w})}{\zeta_i(\mathbf{w}) + \|\mathbf{w}\|^2} \,, \tag{6.20}$$

where $\zeta_i(\mathbf{w})$ is the Hinge loss error for classification directly from the original data $\mathbf{x}$:

$$\zeta_i(\mathbf{w}) = \max(0, 1 + (\max_{\mathbf{h}} E(y, \mathbf{h}|\mathbf{x}_i; \mathbf{w}) - \max_{y' \neq y, \mathbf{h}} E(y', \mathbf{h}|\mathbf{x}_i; \mathbf{w}))) \,. \tag{6.21}$$

6.2.5 Inference

Having computed the optimal parameters $\mathbf{w}^*$ in the training step, our goal is to estimate the optimal label configuration over the testing input, where the optimality is expressed in terms of a cost function. To this end, we maximize the posterior probability and marginalize over the latent variables $\mathbf{h}$ and the privileged information $\mathbf{x}^*$:

$$\begin{aligned} y &= \arg\max_y p(y|\mathbf{x}; \mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y, \mathbf{h}, \mathbf{x}^*|\mathbf{x}; \mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) p(\mathbf{x}^*|\mathbf{x}) \,. \end{aligned} \tag{6.22}$$

In the general case, the training samples $\mathbf{x}$ and $\mathbf{x}^*$ may be considered to be jointly Gaussian, thus the conditional distribution $p(\mathbf{x}^*|\mathbf{x})$ is also Gaussian. We quantized the continuous space of features to a large number of discrete values to approximate the true value of the marginalization of Eq. (6.22). However, to efficiently cope with outlying measurements about the training data, we consider that the training samples $\mathbf{x}$ and $\mathbf{x}^*$ jointly follow a Student's t-distribution. Therefore, the conditional density function $p(\mathbf{x}^*|\mathbf{x})$ is also a Student's t-distribution $\mathrm{St}(\mathbf{x}^*|\mathbf{x}; \mu^*, \Sigma^*, \nu^*)$, where $\mathbf{x}^*$ forms the first $M_{\mathbf{x}^*}$ components of $(\mathbf{x}^*, \mathbf{x})^T$, $\mathbf{x}$ comprises the remaining $M - M_{\mathbf{x}^*}$ components, μ^* is the mean vector, Σ^* is the covariance matrix and $\nu^* \in [0, \infty)$ corresponds to the degrees of freedom of the distribution [321]. Note that by letting the degrees of freedom ν^* go to infinity, we can recover the Gaussian distribution with the same parameters. If the data contain outliers, the degrees of freedom parameter ν^* is weak and the mean and covariance of the data are appropriately weighted in order not to take into account the outliers. More details on how the parameters of the conditional Student's t-distribution $p(\mathbf{x}^*|\mathbf{x})$ are estimated can be found in Appendix A.

Although both distributions $p(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ and $p(\mathbf{x}^*|\mathbf{x})$ belong to the exponential family, the graph in Fig. 6.2 is cyclic, and therefore an exact solution to Eq. (6.22) is generally intractable. For this reason, approximate inference is employed for estimation of the marginal probability by applying the LBP algorithm [303].

6.2.6 Mapping of Discrete Features to Continuous Space

Our model is able to learn the relationship between the input data and the semantic features. Directly comparing the semantic attributes with the raw data is not well principled, as semantic attributes are binary while raw data are not. To this end, we jointly calibrate the different modalities by learning a multiple output linear regression model [152]. Let $\mathbf{x} \in \mathbb{R}^{M \times d}$ be the input raw data and $\mathbf{a} \in \mathbb{R}^{M \times p}$ be the set of semantic attributes. Our goal is to find a set of weights $\boldsymbol{\gamma} \in \mathbb{R}^{d \times p}$, which relates the attributes to the raw features by minimizing a distance function across the input samples and their attributes:

$$\arg\min_{\gamma} \|\mathbf{x}\boldsymbol{\gamma} - \mathbf{a}\|^2 + \eta\|\boldsymbol{\gamma}\|^2 , \tag{6.23}$$

where $\|\boldsymbol{\gamma}\|^2$ is a regularization term and η controls the degree of the regularization. The regularization parameter η was chosen to give the best solution by using a cross validation scheme with $\eta \in [10^{-4}, 1]$. Following a constrained least squares (CLS) optimization problem and minimizing $\|\boldsymbol{\gamma}\|^2$ subject to $\mathbf{x}\boldsymbol{\gamma} = \mathbf{a}$, then Eq. (6.23) has a closed form solution $\boldsymbol{\gamma} = \left(\mathbf{x}^\top\mathbf{x} + \eta I\right)^{-1}\mathbf{x}^\top\mathbf{a}$, where I is the identity matrix. Note that solving this minimization problem is fast since the number of attributes is relatively low, and needs to be solved only once during training.

One drawback of combining features of different modalities is the different frame rate that each modality may have. Thus, instead of directly combining audio and visual features together, we used canonical correlation analysis (CCA) [228] to better exploit the correlation between the different modalities by projecting them onto a common subspace.

6.3 Experimental Results

We evaluated our method on four publicly available datasets in challenging human activity recognition problems. Three different types of privileged information were used: audio signal, human pose, and semantic attribute annotation, and we compared our method with the state-of-the-art.

6.3.1 Datasets

Parliament [5]: This dataset contains 228 video sequences of political speeches, belonging in three behavioral categories: friendly, aggressive, and neutral. It is described in detail in Chapter 4. Figure 5.3 depicts some representative frames of the *Parliament* dataset.

(a) (b) (c)

Figure 6.3: Sample frames from the proposed *Parliament* dataset. (a) Friendly, (b) Aggressive, and (c) Neutral.

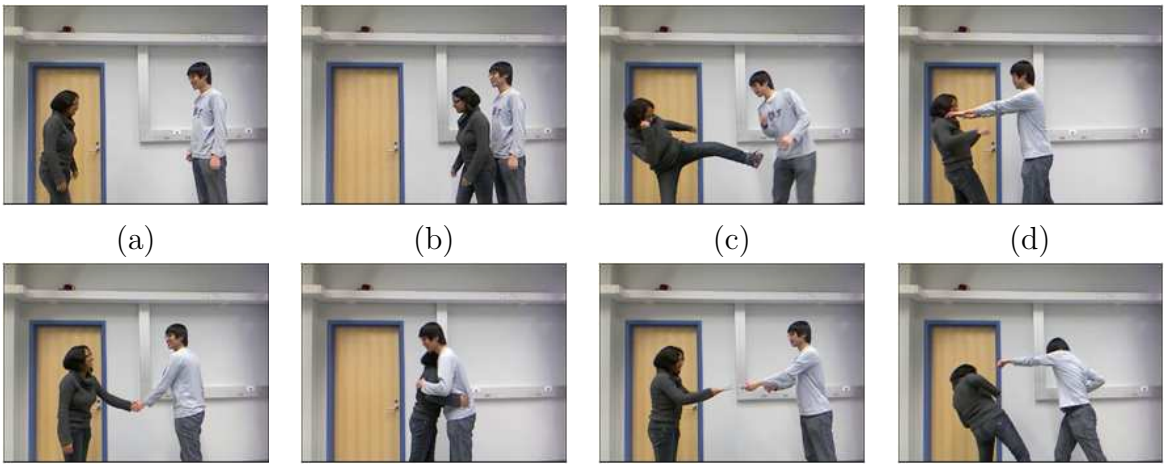
TV human interaction (TVHI) [6]: The TVHI, is a group of 300 video sequences collected by different TV shows and contain four kinds of interactions: *high fives*, *hugs* and *kisses*. This dataset was also used and described in Chapter 5. Some representative frames of the TVHI dataset are illustrated in Figure 6.4.

(a) (b) (c) (d)

Figure 6.4: Sample frames from the TVHI dataset. (a) Hand shake, (b) High five, (c) Hug, and (d) Kiss.

Two-person interaction (TPI) [7]: This dataset is a collection of approximately 300 video sequences depicting two-person interactions captured by a Microsoft Kinect sensor. The dataset contains eight different interaction classes including *approaching*, *departing*, *kicking*, *pushing*, *shaking hands*, *hugging*, *exchanging objects*, and *punching*, which are performed by seven different participants. It also contains three-dimensional

coordinates of 15 joints for each person at each frame. Figure 6.5 shows some sample frames for this dataset.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6.5: Sample frames from the TPI dataset. (a) Approach, (b) Depart, (c) Kick, (d) Push, (e) Shake hands, (f) Hug, (g) Exchange objects, and (h) Punch.

Unstructured social activity attribute (USAA) [8]: The USAA dataset includes eight different semantic class videos of social occasions such as *birthday party*, *graduation party*, *music performance*, *non-music performance*, *parade*, *wedding ceremony*, *wedding dance*, and *wedding reception*. It contains around 100 videos per class for training and testing. Each video is annotated with 69 attributes, which can be broken down into five broad classes: actions, objects, scenes, sounds, and camera movement. Figure 6.6 depicts some representative frames of the USAA dataset.

6.3.2 Feature Selection

For the evaluation of our method on all datasets, we used spatio-temporal interest points (STIP) [97] as our base video representation. First, we extracted local space-time features at a rate of 25 fps using a 72-dimensional vector of HoG and 90-dimensional vector of HoF feature descriptors [241] for each STIP, which captures the human motion between frames. These features were selected because they can capture salient visual motion patterns in an efficient and compact way. In addition, for the TVHI dataset, we also used the provided annotations, which are related to the locations of the persons in each video clip, including the bounding boxes that contain them, the head orientations of each subject in the clips, the pair of subjects who interact with each other, and the corresponding labels. For our experiments on this dataset, we used audio features as privileged information. More specifically, we employed the mel-frequency cepstral coefficients (MFCC) [308] features and their first and second order derivatives. The audio signal was sampled at 16 KHz and processed over 10 *ms* using a Hamming window with 25% overlap. The audio feature

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6.6: Sample frames from the USAA dataset. (a) Birthday party, (b) Graduation party, (c) Music performance, (d) Non-music performance, (e) Parade, (f) Wedding ceremony, (g) Wedding dance, and (h) Wedding reception.

vector consisted of a collection of 13 MFCC coefficients along with the first and second derivatives forming a 39 dimensional audio feature vector.

Furthermore, for the TPI dataset, we used the poses provided by the dataset as privileged information. In particular, along with the positions of the locations of the joints for each person in each frame, we used six more feature types concerning joint distance, joint motion, plane, normal plane, velocity, and normal velocity as described by Yun *et al.* [7]. As basic representation of the video data, we used the STIP features.

Finally, we used the USAA dataset and the provided attribute annotation as privileged information to characterize each class not with an individual label, but with a feature vector of semantic attributes. As a representation of the video data, we used the provided low-level features, which correspond to SIFT [322], STIP, and MFCC features. Table 6.1 summarizes all forms of features used either as regular or privileged for each dataset in our algorithm during training and testing.

6.3.3 Model Selection

The optimal number of hidden states of the model in Fig. 6.2 was estimated based on cross validation, varying the number of hidden states from 3 to 20. The L_2 regularization scale term σ for the non-adaptive methods was set to 10^k, with $k \in \{-3, \ldots, 3\}$. Finally, our model was trained with a maximum of 400 iterations for the termination of the LBFGS optimization method.

The evaluation of our method was performed using 5-fold cross validation to split the datasets into training and test sets, according to the documentation described in each dataset, and we report here the average results over all the examined configurations.

Four variants of our approach are proposed, called *Maximum Likelihood LUPI Hidden Conditional Random Field (ml-HCRF+)*, *Adaptive Maximum Likelihood LUPI Hidden*

Table 6.1: Types of features used for human activity recognition for each dataset. The numbers in parentheses indicate the dimension of the features. The checkmark corresponds to the usage of the specific information as regular or privileged. Privileged features are used only during training.

Dataset	Features (dimension)	Regular	Privileged
Parliament [5]	STIP (162)	✓	
	MFCC (39)		✓
TVHI [6]	STIP (162)	✓	
	Head orientations (2)	✓	
	MFCC (39)		✓
TPI [7]	STIP (162)	✓	
	Pose (15)		✓
USAA [8]	STIP (162)	✓	
	SIFT (128)	✓	
	MFCC (39)	✓	
	Attributes (69)		✓

Conditional Random Field (aml-HCRF+), *Maximum Margin LUPI Hidden Conditional Random Field (mm-HCRF+)*, and *Adaptive Maximum Margin LUPI Hidden Conditional Random Field (amm-HCRF+)*, depending on which learning method we apply (i.e., maximum likelihood or max-margin) and wether we automatically estimate the regularization parameters of the corresponding loss function or not.

6.3.4 Evaluation of Privileged Information

The classification accuracy with respect to the number of hidden states is depicted in Fig. 6.7. We may observe that all four variants have a similar behavior as the number of hidden states increases. The max-margin HCRF+ approach seems to perform better than the maximum likelihood HCRF+ approach for all datasets. Moreover, the performance of the adaptive methods is equally good in many cases they perform higher than the non-adaptive methods HCRF+ variants. The optimal number of hidden states for the Parliament dataset is eight for both the non-adaptive approaches and 11 for the adaptive approaches, respectively. The best accuracy for the TVHI dataset is seven, when the maximum likelihood HCRF+ method is used, while the max-margin HCRF+ variant requires more hidden states (12) to reach the maximum accuracy. On the other hand, the adaptive methods may achieve their highest performance with less hidden states. The maximum accuracy for the TPI dataset is reached for 11 hidden states for the proposed maximum likelihood HCRF+ model and for 14 hidden states for its adaptive counterpart. The max-margin HCRF+ approach requires at least 16 hidden states to achieve the highest accuracy and its adaptive form requires only 11. Finally, the USAA dataset achieves its best accuracy for both the non-adaptive approaches in 15 hidden states their

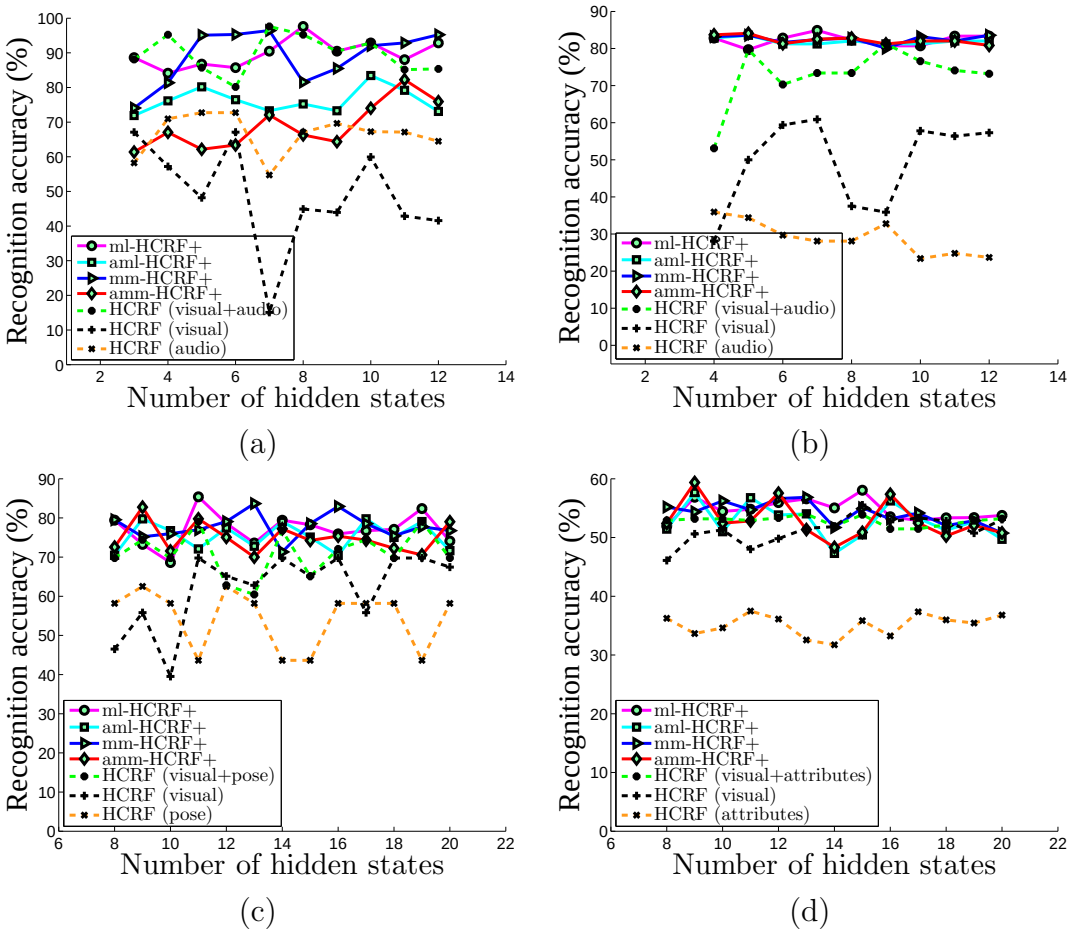


Figure 6.7: Comparison of the recognition accuracy of the four different variants of the proposed method and standard HCRF model with respect to the number of hidden states for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7], and (d) the USAA [8] datasets. The text in parentheses in the legend of each figure corresponds to the type of information used both for training and testing.

adaptive counterpart methods require 11 and nine hidden states, respectively.

In Fig. 6.7 we may observe that the standard HCRF model suffers from large fluctuations in recognition accuracy as the number of hidden states increases. This is because the number of hidden states plays a crucial role in the recognition process. Many hidden states may lead to model overfitting, while few hidden states may cause underfitting. This would be resolved by the estimation of the optimal number of hidden states during learning, but this is not straightforward for this model. For example, when only the visual information is used for both training and testing, we may see that there exist very large variations in the recognition accuracy for all datasets and for few hidden states, as this number may be small and the model may not generalize under such poor conditions. We may also observe that for these datasets the performance of each modality alone is kept significantly lower for all configurations of hidden states, which reinforces the fact that privileged information may help to construct better classification models.

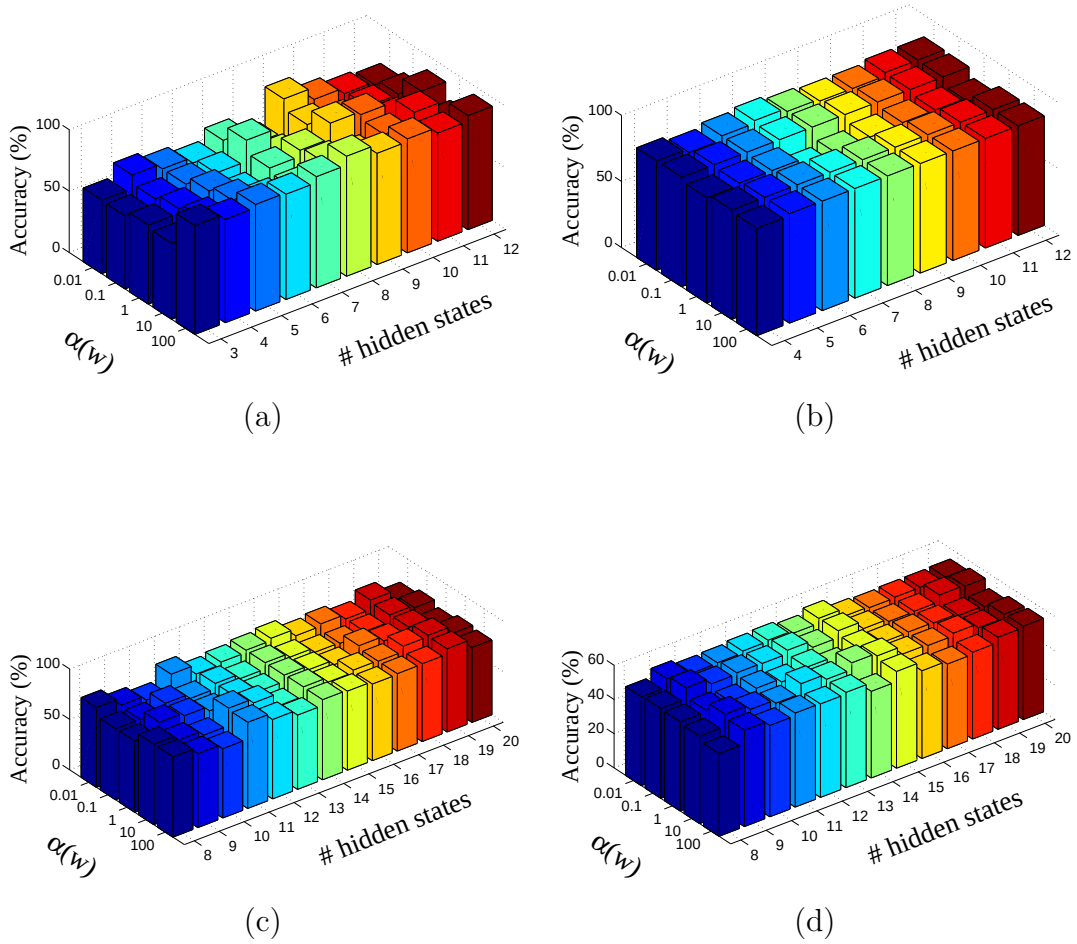


Figure 6.8: Recognition performance of the proposed maximum likelihood variant as function of the regularization parameter and the number of hidden states for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7] and (d) the USAA [8] datasets.

The behavior of the proposed adaptive model as a function of the regularization parameters and the number of hidden states for all four datasets for the aml-HCRF+ and amm-HCRF+ is depicted in Fig. 6.8 and Fig. 6.9, respectively. To be consistent to the non-adaptive methods, the real-valued regularization parameters were quantized from the continuous to the discrete space with $\alpha(\mathbf{w}) = 10^k, k \in \{-2, \ldots, 2\}$ and the results were averaged. We may observe that the behavior of the recognition accuracy is smooth for the different values of $\alpha(\mathbf{w})$ and the number of hidden states, which indicates that the automatic estimation of $\alpha(\mathbf{w})$ is robust and may lead to high classification accuracies with performances close to the non-adaptive approaches.

6.3.5 Comparison of Learning Frameworks

We compare the results of our method with several state-of-the-art methods. In particular, to show the benefit of using robust privileged information we compared our method both with state-of-the-art methods with and without incorporating the LUPI paradigm.

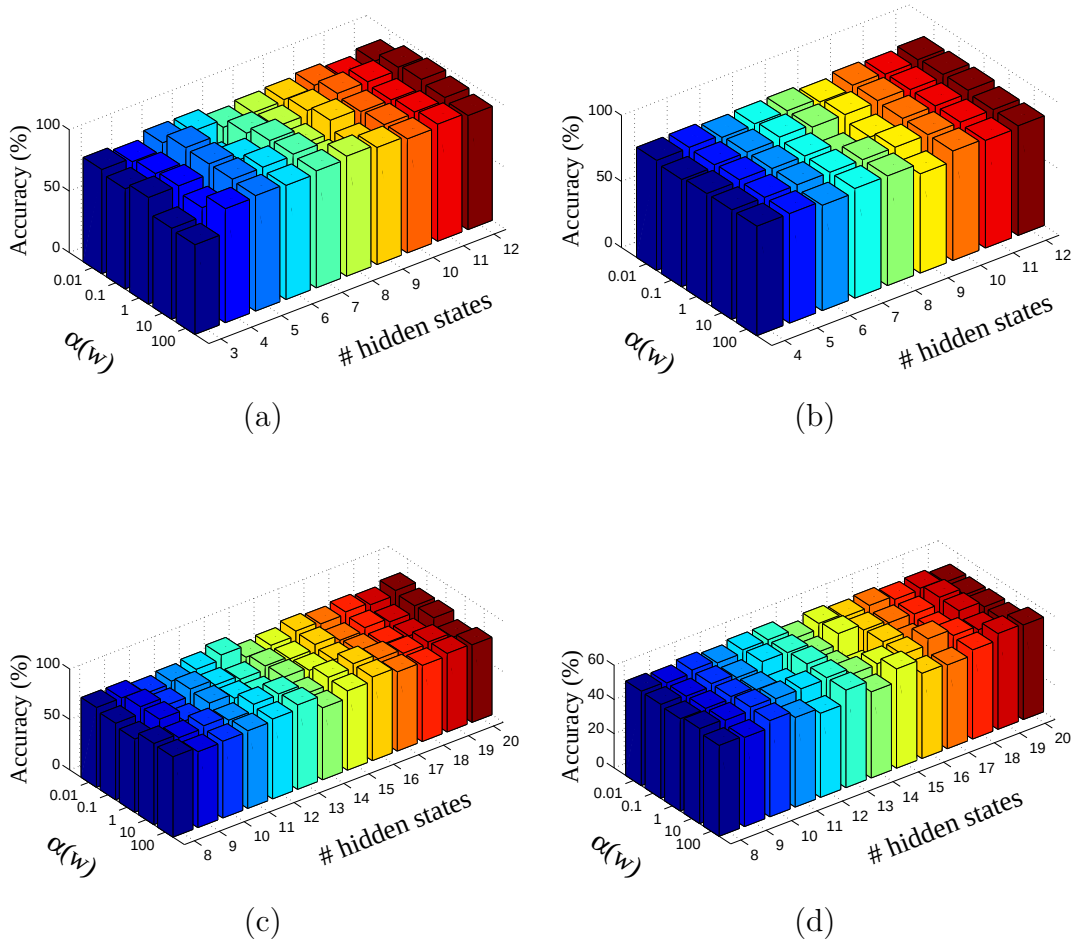


Figure 6.9: Recognition performance of the proposed max-margin variant as function of the regularization parameter and the number of hidden states for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7] and (d) the USAA [8] datasets.

The first chronologically method that integrated the LUPI framework for classification purposes was SVM+ [313]. In a nutshell, SVM+ consists of optimizing the hyperplane parameters such that it can minimize the probability of incorrect classifications and increase the convergence rate. A brief description of SVM+ can be found in Appendix B. Also, to demonstrate the efficacy of the robust privileged information to the problem of human activity recognition and show how it can be used for constructing accurate classifiers, we compared it with ordinary SVM and HCRF, as if they could access both the original and the privileged information at test time. This means that we do not differentiate between regular and privileged information, but use both forms of information as regular to infer the underlying class label instead. Moreover, to complete the study, we also trained an HCRF model that uses only the regular and only the privileged information for training and testing. To distinguish between the different types of information that the HCRF model may use, we specifically report the type of feature in parentheses after the HCRF caption. Furthermore, for the SVM+ and SVM we consider a one-versus-one

Table 6.2: Comparison of the classification accuracies (%) on Parliament dataset [5].

Method	Overall	Aggressive	Friendly	Neutral
Methods without privileged information				
Vrigkas *et al.* [5]	85.5	100.0	60.7	95.8
Wang and Schmid [323]	66.6	67.9	60.0	71.1
SVM [25]	72.6	76.9	69.8	71.1
HCRF (visual+audio) [27]	**97.6**	92.7	100.0	100.0
HCRF (visual) [27]	67.1	50.0	66.7	84.6
HCRF (audio) [27]	72.7	85.7	55.6	76.9
Methods with privileged information				
Wang and Ji [316]	59.2	77.9	39.2	60.5
Sharmanska *et al.* [317]	57.7	57.1	58.1	57.8
Wang *et al.* [324]	96.9	90.7	100.0	100.0
SVM+ [313]	78.4	77.5	68.9	88.7
ml-HCRF+	**97.6**	92.9	100.0	100.0
aml-HCRF+	83.5	85.7	80.0	84.6
mm-HCRF+	96.5	92.6	100.0	97.4
amm-HCRF+	82.3	85.7	61.1	100.0

decomposition of multi-class classification scheme and average the results for every possible configuration. Finally, the optimal parameters for the SVM and SVM+ were selected using cross validation.

Table 6.2 compares the proposed approach with state-of-the-art methods on the human activity classification task on the Parliament dataset. The proposed maximum likelihood HCRF+ method has highest recognition accuracy (97.6%) among the other variants of the proposed model, while it achieves the same accuracy with the standard HCRF model. Although both the adaptive HCRF+ approaches may perform worse than the non adaptive variants, they can still achieve better results than the majority of the state-of-the-art methods. The estimation of the regularization parameters for the adaptive variants of the proposed method depends on the input features. Features that belong to the background may influence the estimation of the regularization parameters as they may serve as background noise.

It is also worth mentioning that our method is able to increase the recognition accuracy by nearly 38% with respect to the methods of Wang and and Ji [316] and the method of Sharmanska *et al.* [317], which also incorporate the LUPI paradigm. This significantly high increase in recognition accuracy indicates the strength of the proposed method. Moreover, the performance of the proposed approach on the Parliament dataset is higher approximately by 19% than the SVM+ model and 25% than the standard SVM approach. The Parliament dataset contains large intra-class variabilities. For example the interaction between an arm lift and the raise in the voice may not exclusively be combined together

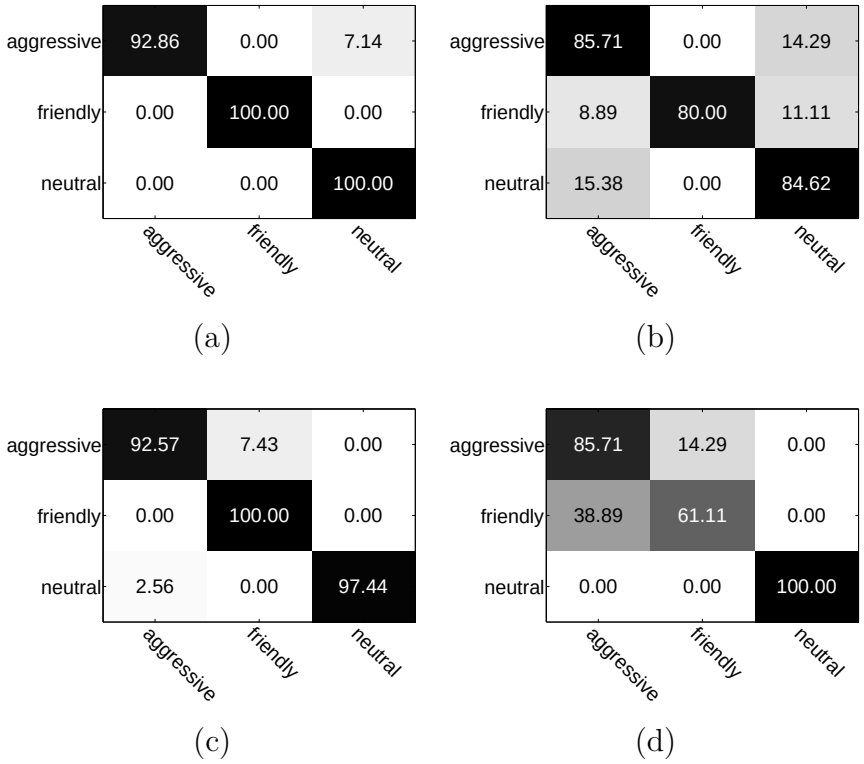


Figure 6.10: Confusion matrices for the classification results of the proposed HCRF+ approach for the *Parliament* dataset [5] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants.

as some features may act as outliers and affect the classification accuracy.

The resulting confusion matrices of the proposed method for the optimal number of hidden states for the Parliament dataset are depicted in Figure 6.10. It is worth mentioning that for this dataset the classification errors between different classes are relatively small, while the maximum likelihood HCRF+ approach may perfectly recognize the classes *friendly* and *neutral*.

Table 6.3 demonstrates the classification results on the TVHI dataset. For this dataset, we significantly managed to increase the classification accuracy by approximately 10%, with respect to the LUPI-based SVM+ and Wang and Ji [316] approaches, as our approach achieves very high recognition accuracy for this dataset (84.9%). The improvement of our method compared to the method of Sharmanska *et al.* [317] was even higher. On the other hand, the method of Wang *et al.* [324] was able to yield similar results to our ml-HCRF+ approach, as it achieved an accuracy of 84.4%. It is also worth mentioning that when our method is compared with methods that do not use privileged information, it is able to increase the recognition accuracy. Also, if only privileged information (HCRF (audio)) is used as regular features for classification, the recognition accuracy is notably lower than when using visual information (HCRF (visual)) for the classification task. In general, when privileged information alone is used as regular it may not be suffiecient for correct classification of an action into its respective category, since finding proper

Table 6.3: Comparison of the classification accuracies (%) on TVHI dataset [6].

Method	Overall	Hand Shake	High Five	Hug	Kiss
Methods without privileged information					
Patron-Perez *et al.* [6]	64.2	57.8	51.1	71.2	76.5
Hoai and Zisserman [251]	56.3	55.8	60.2	60.8	48.2
Marín-Jiménez *et al.* [58]	54.5	36.3	59.4	66.9	40.9
Wang and Schmid [323]	76.1	76.2	74.6	74.8	74.6
SVM [25]	75.9	74.6	76.3	75.8	76.3
HCRF (visual+audio) [27]	81.3	87.5	56.3	87.5	93.8
HCRF (visual) [27]	60.9	56.3	25.0	87.5	75.0
HCRF (audio) [27]	35.9	12.5	12.5	43.8	75.0
Methods with privileged information					
Wang and Ji [316]	74.8	74.6	76.3	72.2	76.3
Sharmanska *et al.* [317]	65.2	78.3	54.8	74.3	53.5
Wang *et al.* [324]	84.4	93.8	81.2	75.1	87.5
SVM+ [313]	75.0	74.6	76.3	72.8	76.2
ml-HCRF+	**84.9**	97.2	81.3	72.9	87.5
aml-HCRF+	83.6	93.8	81.3	71.8	87.5
mm-HCRF+	83.6	93.8	81.3	72.5	87.5
amm-HCRF+	82.9	93.8	81.3	68.8	87.5

privileged information is not always a straightforward process.

Figure 6.11 illustrates the confusion matrices of four variant of the proposed method for the TVHI dataset. The maximum likelihood HCRF+ and the max-margin HCRF+ have the smallest classification errors. The category *hand shake* is the most commonly confused class as the remaining three classes have many false positives for this class. This is due to the fact that the TVHI dataset has large intra-class variability.

The classification accuracies for the TPI dataset are reported in Table 6.4. The best accuracy was achieved by the ml-HCRF+ approach, where we were able to improve the accuracy by nearly 12% with respect to the method of Yun *et al.* [7], while compared to the standard HCRF model, we were better by nearly 4%. Comparing our method to methods that do not use privileged information we increased the classification accuracy in all cases. An interesting characteristic of the non-privileged methods HCRF (visual) and HCRF (audio) is that despite the fact that for some classes these methods were able to perfectly recognize the underlying activity, they failed to recognize some of the classes as the rate of false positives may reach 100%. This observation, reinforces the intuition that different modalities may help in constructing better classifiers. Considerably high improvements are also reported when comparing our methods with state-of-the-art methods that employ privileged information. Closer to our results were the method of Wang *et al.* [324]. We may also observe that all four variants outperform all privileged and non-privileged based

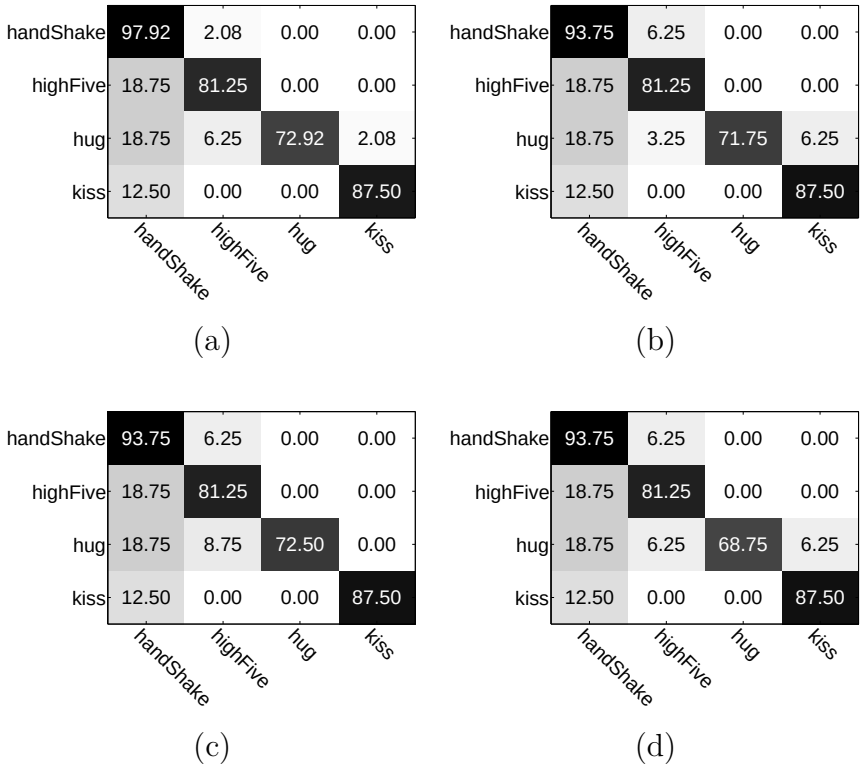


Figure 6.11: Confusion matrices for the classification results of the proposed HCRF+ approach for the TVHI dataset [6] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants.

Table 6.4: Comparison of the classification accuracies (%) on TPI dataset [7].

Method	Overall	Approach	Depart	Kick	Push	Shake Hands	Hug	Exchange Objects	Punch
Methods without privileged information									
Yun *et al.* [7]	73.8	88.0	96.0	71.0	69.0	69.0	50.0	79.0	63.0
Wang and Schmid [323]	79.6	76.2	74.6	78.6	78.9	81.4	79.2	84.3	83.5
SVM [25]	79.4	74.9	67.2	68.7	76.9	100.0	59.4	89.4	100.0
HCRF (visual+pose) [27]	81.4	100.0	33.3	100.0	66.7	66.7	75.0	100.0	83.3
HCRF (visual) [27]	69.8	100.0	100.0	100.0	66.7	100.0	0.0	100.0	0.0
HCRF (pose) [27]	62.5	100.0	0.0	100.0	100.0	0.0	0.0	100.0	100.0
Methods with privileged information									
Wang and Ji [316]	62.4	79.5	61.4	59.2	60.0	59.7	60.5	56.4	62.6
Sharmanska *et al.* [317]	56.3	51.6	79.2	40.9	60.0	74.1	39.9	43.6	61.2
Wang *et al.* [324]	83.7	100.0	66.7	75.0	66.7	66.7	75.5	100.0	100.0
SVM+ [313]	79.4	76.4	72.6	73.2	91.5	70.2	73.2	81.4	100.0
ml-HCRF+	**85.4**	100.0	83.3	100.0	100.0	66.7	33.3	100.0	100.0
aml-HCRF+	79.8	100.0	100.0	75.0	77.8	100.0	50.0	66.7	66.7
mm-HCRF+	83.7	100.0	75.0	100.0	100.0	66.7	25.0	100.0	100.0
amm-HCRF+	82.8	100.0	66.7	83.4	66.7	66.7	75.0	100.0	100.0

methods.

The confusion matrices for the TPI dataset are depicted in Figure 6.12. Note that the classification error is relatively small, as only a few classes are confused with each other (e.g., the class *hugging* may be confused with the class *hand shaking*), in all four variants. For both the maximum likelihood and the the max-margin HCRF+ approaches, five out

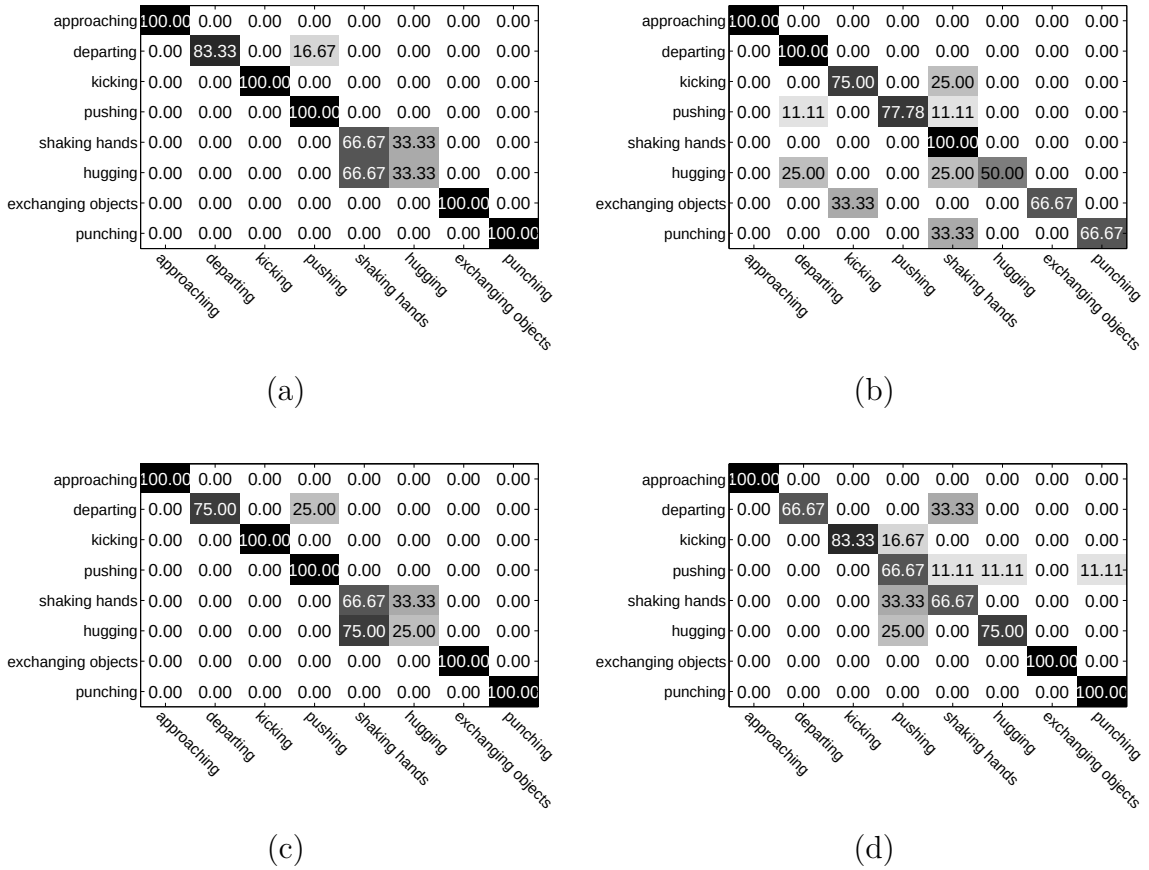


Figure 6.12: Confusion matrices for the classification results of the proposed HCRF+ approach for the TPI dataset [7] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants.

Table 6.5: Comparison of the classification accuracies (%) on USAA dataset [8].

Method	Overall	Birthday	Graduation	Music	Non-music	Parade	Ceremony	Dance	Reception
Methods without privileged information									
Wang and Schmid [323]	55.6	52.8	55.3	57.1	58.3	60.2	49.7	59.6	40.1
SVM [25]	47.4	47.5	47.9	49.4	45.7	48.7	38.2	36.5	45.9
HCRF (visual+attributes)[27]	54.0	79.8	59.6	48.5	68.3	61.5	4.4	69.8	21.2
HCRF (visual) [27]	55.5	74.8	50.5	76.4	50.5	79.1	4.3	80.2	19.2
HCRF (attributes) [27]	37.4	22.2	41.4	63.7	47.5	35.2	14.1	56.3	0.0
Methods with privileged information									
Wang and Ji [316]	48.5	32.9	44.6	52.7	48.9	52.0	49.4	54.7	53.0
Sharmanska *et al.* [317]	56.3	56.9	47.8	62.0	62.6	67.1	51.8	57.5	44.4
Wang *et al.* [324]	55.3	58.6	68.7	58.4	67.3	74.7	17.4	75.0	15.4
SVM+ [313]	48.5	52.7	49.9	53.3	50.9	51.6	48.7	41.1	32.5
ml-HCRF+	58.1	78.8	59.6	74.3	60.4	70.3	11.3	87.5	23.5
aml-HCRF+	57.5	78.8	57.6	78.2	70.3	67.0	3.3	78.1	23.1
mm-HCRF+	56.8	79.8	63.6	79.2	59.4	54.9	14.6	85.4	17.5
amm-HCRF+	**59.4**	78.8	61.6	77.2	69.3	69.2	18.3	79.2	21.2

of the eight classes were perfectly recognized. Accordingly, both the adaptive maximum likelihood and max-margin HCRF+ methods have also performed remarkably well and were able to perfectly recognize three out of eight categories.

The classification results for the USAA dataset are summarized in Table 6.5. The

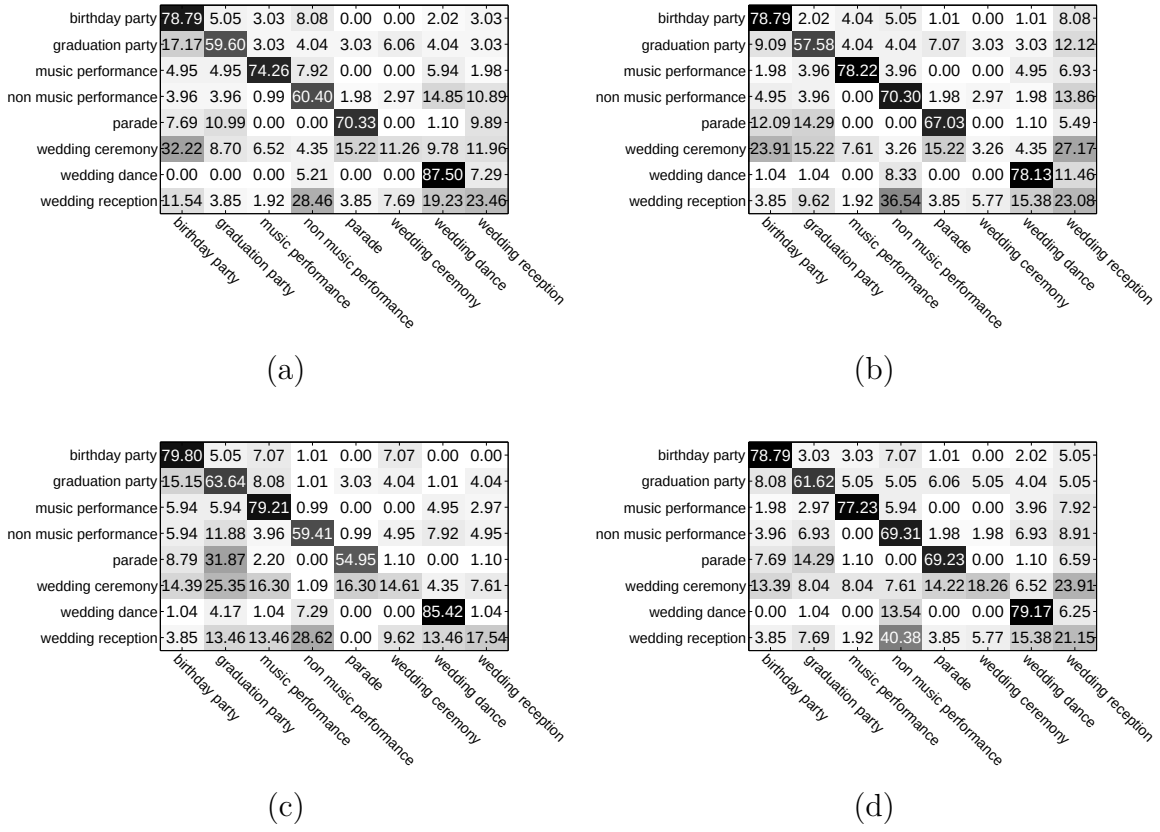


Figure 6.13: Confusion matrices for the classification results of the proposed HCRF+ approach for the USAA dataset [8] for (a) the ml-HCRF+, (b) the aml-HCRF+, (c) the mm-HCRF+, and (d) the amm-HCRF+ variants.

combination of both raw data and attribute representation of human activities on the USAA dataset significantly outperforms the SVM+ baseline and the method of Wang and Ji [316] by increasing the classification accuracy by approximately 11% for the amm-HCFR+ model. An improvement of 3% with respect to the methods of Sharmanska *et al.* [317] and Wang *et al.* [324] was also achieved. Furthermore, the adaptive variants of the proposed method perform better than their non-adaptive counterparts for this dataset. Automatic estimation of the regularization parameters provides more flexibility to the model as it allows the model to adjust its behavior according to the training data.

In general, our method is able to robustly use privileged information in a more efficient way than the SVM+ and the other LUPI based methods, by exploiting the hidden dynamics between the video clips and the privileged information. We can also observe that the proposed method outperforms both the SVM and HCRF models. Note that the HCRF (attributes) approach shows the lowest results among all other methods as the use of binary features for training and testing may contain inherent biases and thus model cannot generalize under unknown video sequences. However, the combination of visual and semantic features does not suffer from the biasing problem due to feature calibration and their projection to a common subspace using Eq. (6.23).

Figure 6.13 depicts the confusion matrices for the USAA dataset. It is interesting to

Table 6.6: p-values of the proposed method for the *Parliament* dataset [5].

Method	**MLHCRF+**	**AMLHCRF+**	**MMHCRF+**	**AMMHCRF+**
Vrigkas *et al.* [5]	0.0154	0.5978	0.1759	0.5610
Wang and Schmid [323]	0.0011	0.0001	0.0015	0.0345
SVM [25]	0.0022	0.0001	0.0022	0.0149
HCRF [27]	0.1283	0.9883	0.6074	0.9055
HCRF (visual) [27]	0.0234	0.0429	0.0270	0.0440
HCRF (audio) [27]	0.0064	0.0465	0.0089	0.0448
Wang and Ji [316]	0.0131	0.0184	0.0142	0.0189
Sharmanska *et al.* [317]	0.0005	0.0001	0.0004	0.0282
Wang *et al.* [324]	0.0128	0.1988	0.4613	0.5955
SVM+ [313]	0.0102	0.0116	0.0152	0.0201

observe that for this dataset the different classes may be strongly confused. For example, the class *wedding ceremony* is confused with the class *birthday party* and the class *wedding reception* is confused with the class *non-music performance* as the dataset has large intra-class variabilities, while the corresponding classes may share the same attribute representation as different videos may have been captured under similar conditions.

The main strength of the proposed method is that it achieves remarkably good classification results, when the LUPI framework is incorporated with the standard HCRF model. The probabilistic approach of privileged learning and the automatic estimation of the regularization parameters provide flexibility to the model, which constitute an important cue for high classification performance. The performance of the adaptive based methods is close to the non-adaptive ones and in many cases it is higher. Moreover, as the regularization parameters are estimated during training from the training examples, it is not necessary to re-estimate these parameters for a new problem, which reduces the required time for performing cross-validation on the data.

In order to provide a statistical evidence of the recognition accuracy, we computed the p-values of the obtained results with respect to the compared methods. The null hypothesis was defined as: the mean performances of the proposed model are the same as those of the state-of-the-art methods; and the alternative hypothesis was defined as: the mean performances of the proposed model are higher than those of the state-of-the-art methods. Paired t-tests showed that the results were statistically significant for all datasets.

For the Parliament dataset (Table 6.6), we may observe that for the majority of the comparisons all four variant of the proposed approach reject the null hypothesis as all values are greater than the critical value (95% of significance level). When the proposed method is compared to the HCRF model, which uses both audio and visual modalities for training and testing, the p-values are grater than the threshold of 0.05. However, this does not mean that our results were achieved due to chance as both the proposed method and the HCRF model may yield comparable results. The p-values for the TVHI dataset are reported in Table 6.7. The null hypothesis is rejected for the majority of the

Table 6.7: p-values of the proposed method for the TVHI dataset [6].

Method	**MLHCRF+**	**AMLHCRF+**	**MMHCRF+**	**AMMHCRF+**
Patron-Perez *et al.* [6]	0.0185	0.0191	0.0178	0.0252
Hoai and Zisserman [251]	0.0033	0.0034	0.0031	0.0049
Marín-Jiménez *et al.* [58]	0.0127	0.0129	0.0123	0.0157
Wang and Schmid [323]	0.0323	0.0366	0.0320	0.0494
SVM [25]	0.0487	0.0453	0.0486	0.0414
HCRF [27]	0.0319	0.0374	0.0367	0.0407
HCRF (visual) [27]	0.0607	0.0672	0.0650	0.0765
HCRF (audio) [27]	0.0100	0.0098	0.0096	0.0108
Wang and Ji [316]	0.0273	0.0275	0.0241	0.0451
Sharmanska *et al.* [317]	0.0148	0.0201	0.0186	0.0275
Wang *et al.* [324]	0.3540	0.8584	0.8668	0.8812
SVM+ [313]	0.0301	0.0309	0.0273	0.0500

Table 6.8: p-values of the proposed method for the TPI dataset [7].

Method	**MLHCRF+**	**AMLHCRF+**	**MMHCRF+**	**AMMHCRF+**
Yun *et al.* [7]	0.0474	0.0476	0.1044	0.0942
Wang and Schmid [323]	0.2317	0.5020	0.3312	0.2795
SVM [25]	0.2278	0.4992	0.3203	0.3175
HCRF [27]	0.2942	0.4617	0.2942	0.2128
HCRF (visual) [27]	0.0143	0.0216	0.0179	0.0236
HCRF (pose) [27]	0.0327	0.0187	0.0353	0.0870
Wang and Ji [316]	0.0076	0.0058	0.0178	0.0011
Sharmanska *et al.* [317]	0.0051	0.0001	0.0110	0.0068
Wang *et al.* [324]	0.3033	0.5917	0.4116	0.2203
SVM+ [313]	0.2095	0.5106	0.3192	0.2994

cases. That is, for two out of 12 cases the p-values were less than the significance level of 0.05. Therefore, we may conclude that the null hypothesis can be rejected and the improvements obtained by our model are statistically significant.

Table 6.8 presents the statistical significance values for the TPI dataset. We may observe that the null hypotheses is rejected for only half of the cases, while for the rest methods the p-values are greater than the significant threshold. However, the rejection of the null hypothesis does not necessarily indicate that the results are not practical significance. Finally, the statistical significance values between the proposed method and the different state-of-the-art methods for the USAA dataset are shown in Table 6.9. We may see that for almost all cases the null hypothesis is rejected. In general, we may conclude that the null hypothesis can be rejected for and the improvements obtained by our model are statistically significant and not due to chance.

Table 6.9: p-values of the proposed method for the USAA dataset [8].

Method	**MLHCRF+**	**AMLHCRF+**	**MMHCRF+**	**AMMHCRF+**
Wang and Schmid [323]	0.0295	0.0359	0.0368	0.0222
SVM [25]	0.0465	0.0.0399	0.1024	0.0492
HCRF [27]	0.0505	0.0745	0.1353	0.0219
HCRF (visual) [27]	0.0404	0.1167	0.1547	0.0355
HCRF (attributes) [27]	0.0018	0.0029	0.0017	0.0001
Wang and Ji [316]	0.0158	0.0198	0.0200	0.0122
Sharmanska *et al.* [317]	0.3975	0.4575	0.4727	0.3227
Wang *et al.* [324]	0.1719	0.2739	0.3129	0.0810
SVM+ [313]	0.0108	0.0141	0.0137	0.0468

6.4 Conclusion

To address the problem of missing information, a novel probabilistic classification model based on robust learning using a privileged information paradigm, called HCFR+, was presented. The proposed model is made robust using Student's t-distributions to model the conditional distribution of the privileged information. Two variants for training in the LUPI framework were proposed. The first variant uses maximum likelihood (MLHCRF+) and the second uses maximum margin (MMHCFR+) learning. Moreover, the regularization parameters of the loss functions for both maximum likelihood and max-margin approaches were automatically estimated allowing the model to be more flexible.

Using auxiliary information about the input data, the proposed model was able to produce better classification results than the standard HCRF [27] approach by incorporating into the classification model auxiliary information about the input data, which is available only during model training. The performance of the proposed method was evaluated on four publicly available datasets and various forms of data that can be used as privileged were tested. The experimental results indicated that robust privileged information along with the regular input data for training the model ameliorates the recognition performance.

The proposed HCRF+ method and its variants achieved notably higher performance than all the compared classification schemes. In particular, the proposed method is able to flexibly understand multimodal human activities with high accuracy, when not the same amount of information is available during testing. Also, high recognition accuracy with less effort than standard cross validation based classification schemes was achieved by automatically estimating the regularization parameters during learning. Since the combination of multimodal data falls natural to the human perception of understanding complex activities, the incorporation of such information to the proposed model allows us to significantly increase the recognition accuracy for natural video sequences. Furthermore, it was shown that the combination of multimodal data constitute a strong attribute for discriminating between different classes in real-world vision problems, rather than learning each modality separately.

Chapter 7

Active Privileged Learning of Human Activities from Weakly Labeled Samples

7.1 Introduction

7.2 Active Privileged Leaning

7.3 Experimental Results

7.4 Conclusion

7.1 Introduction

Most of the recognition systems including classification systems based on the LUPI paradigm assume that labeled training data are easy to obtain. However, knowing a priori the label of all training examples may not always be feasible for large databases as the cost for manually labeling all samples may be prohibitively large. To address this limitation, active learning has been proposed [325]. The idea of active learning is closely related to semi-supervised learning, where during training, labeled and unlabeled data co-exist. The aim of active learning is to actively select the most informative unlabeled samples according to a specified criterion, query their label and use them as training data to construct a stronger classifier. Active learning has been used with several classification models such as SVM [326], conditional random fields [327] and radial basis function networks [328].

An interesting application of active learning is the automatic annotation of ongoing activities in unsegmented video sequences for detecting and localizing human actions [254]. Hasan and Roy-Chowdhury [329] proposed an incremental algorithm for actively

learning new actions from streaming videos. However, one of the main problems of active learning is how to define an effective criterion for selecting unlabeled samples [330]. To this end, the same authors [331] combined entropy and mutual information to handle inter and intra-relationships between training data through incremental update of the classification model to learn human activities. Finally, Long *et al.* [332] considered an action recognition method that exploits active learning to cope with multiple and noisy labels.

Previous methods can either handle information that is not available during testing, or cope with missing labels during training but cannot address both problems simultaneously. In this chapter, a novel classification method that combines the LUPI paradigm and active learning for identifying human activities in a semi-supervised framework using hidden conditional random fields (HCRFs) [27], called active-HCRF+ (a-HCRF+) is presented. The proposed method exploits privileged information as an additional input during training to learn the conditional probability distribution between human activities and observations. To reduce tedious human effort in data annotation, an incremental pool-based active learning technique is adopted to actively select unlabeled training samples for which the uncertainty about their actual class label is reduced.

7.2 Active Privileged Leaning

We consider a labeled dataset $\mathcal{D} = \{(\mathbf{x}_{i,j}, \mathbf{x}^*_{i,j}, y_i)\}_{i=1}^N$ with N video sequences, where $\mathbf{x}_{i,j} \in \mathbb{R}^{M_\mathbf{x} \times T}$ is an observation sequence of length T with $j = 1 \dots T$, which belongs in feature space $\mathcal{X}$. For example, $\mathbf{x}_{i,j}$ might correspond to the j^{th} frame of the i^{th} video sequence. Furthermore, y_i corresponds to a class label defined in a finite label set $\mathcal{Y}$. Also additional information about the observations $\mathbf{x}_i$ is encoded in a feature vector $\mathbf{x}^*_{i,j} \in \mathbb{R}^{M_{\mathbf{x}^*} \times T}$ and belongs to feature space $\mathcal{X}^*$. This information is provided only at the training step and it is not available during testing, while not any assumption about the form of the privileged data is made. In what follows, we omit indices i and j for simplicity.

7.2.1 a-HCRF+ Model Formulation

The a-HCRF+ model is defined by a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 7.1). The proposed model is a member of the exponential family and the probability of the class label given an observation sequence is given by:

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) &= \sum_{\mathbf{h}} p(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) \\ &= \frac{1}{A(\mathbf{w})} \sum_{\mathbf{h}} \exp\left(E(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w})\right) , \end{aligned} \tag{7.1}$$

where $\mathbf{h} = \{h_1, h_2, \dots, h_T\}$, with $h_j \in \mathcal{H}$ is a set of latent variables and $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters. Finally, $E(y, \mathbf{h}|\mathbf{x}; \mathbf{w})$ is a function of sufficient statistics and

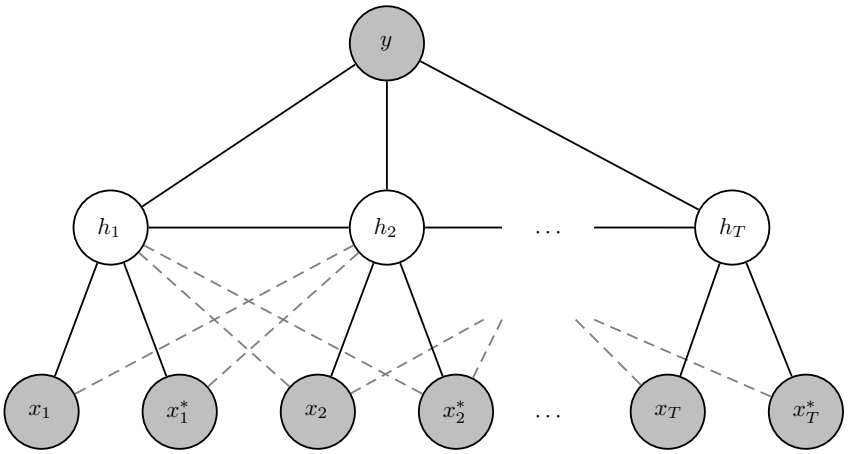


Figure 7.1: Graphical representation of the chain structure model. The grey nodes are the observed features (x_i and x_i^*), and the unknown labels (y). The white nodes are the hidden variables (h).

$A(\mathbf{w})$ is the partition function ensuring normalization:

$$A(\mathbf{w}) = \sum_{y'} \sum_{\mathbf{h}} \exp\left(E(y', \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w})\right) . \tag{7.2}$$

Different sufficient statistics $E(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ in Eq. (7.1) define different distributions. Generally, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$E(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) = \sum_{j \in \mathcal{V}} \Phi(y, h_j, \mathbf{x}_j, \mathbf{x}_j^*; \boldsymbol{\theta}) + \sum_{j,k \in \mathcal{E}} \Psi(y, h_j, h_k; \boldsymbol{\omega}) , \tag{7.3}$$

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned.

The unary potential is expressed by:

$$\Phi(y, h_j, \mathbf{x}_j, \mathbf{x}_j^*; \boldsymbol{\theta}) = \sum_j \sum_\ell \phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) + \sum_j \phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) + \sum_j \phi_3(h_j, \mathbf{x}_j^*; \boldsymbol{\theta}_3) , \tag{7.4}$$

and it can be seen as a state function consisting of three different feature functions. The label feature function models the relationship between the label y and the hidden variables h_j, and it is expressed by:

$$\phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) = \sum_{\lambda \in \mathcal{Y}} \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_{1,\ell} \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) , \tag{7.5}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if its argument is true and 0 otherwise. The observation feature function, which models the relationship between the hidden variables h_j and the observations $\mathbf{x}_j$, is defined by:

$$\phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_2^\top \mathbb{1}(h_j = a) \mathbf{x}_j . \tag{7.6}$$

Finally, the privileged feature function, which models the relationship between the hidden variables h_j and the privileged information $\mathbf{x}_j^*$, is defined by:

$$\phi_3(h_j, \mathbf{x}_j^*; \boldsymbol{\theta}_3) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_3^\top \mathbb{1}(h_j = a)\mathbf{x}_j^* \,. \tag{7.7}$$

The pairwise potential is expressed by:

$$\Psi(y, h_j, h_k; \boldsymbol{\omega}) = \sum_{\substack{\lambda \in \mathcal{Y} \\ a,b \in \mathcal{H}}} \sum_{\ell} \boldsymbol{\omega}_\ell \mathbb{1}(y = \lambda)\mathbb{1}(h_j = a)\mathbb{1}(h_k = b) \,. \tag{7.8}$$

It is a transition function and represents the association between a pair of connected hidden states h_j and h_k and the label y.

7.2.2 Learning and Inference

In the training step, the optimal parameters $\mathbf{w}^*$ are estimated by maximizing the following loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \frac{1}{2\sigma^2}\|\mathbf{w}\|^2 \,. \tag{7.9}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ and quantifies how well the distribution in Eq. (7.1) defined by the parameter vector $\mathbf{w}$ matches the labels y.

The second term is a L_2 regularization Gaussian prior with variance σ^2. The use of hidden variables makes the optimization of Eq. (7.9) non-convex, thus, a global solution is not guaranteed and we can estimate $\mathbf{w}^*$ that are locally optimal. The loss function is optimized using the limited-memory BFGS (LBFGS) method [306] to minimize the negative log-likelihood of the data.

Having computed the optimal parameters $\mathbf{w}^*$ in the training step, our goal is to estimate the optimal label configuration over the testing input. We maximize the posterior probability and marginalize over the latent variables $\mathbf{h}$ and the privileged information $\mathbf{x}^*$:

$$\begin{aligned} y &= \arg\max_y p(y|\mathbf{x}; \mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y, \mathbf{h}, \mathbf{x}^*|\mathbf{x}; \mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) p(\mathbf{x}^*|\mathbf{x}) \,. \end{aligned} \tag{7.10}$$

In the general case, the training samples $\mathbf{x}$ and $\mathbf{x}^*$ may be considered to be jointly Gaussian, thus the conditional distribution $p(\mathbf{x}^*|\mathbf{x})$ is also a Gaussian distribution. We quantized the continuous space of features to a large number of discrete values to approximate the true value of the marginalization of Eq. (7.10). To efficiently cope with outlying measurements about the training data, we consider that the training samples $\mathbf{x}$ and $\mathbf{x}^*$ jointly follow a Student's t-distribution. More details on how the parameters of the conditional Student's t-distribution $p(\mathbf{x}^*|\mathbf{x})$ are estimated can be found in Appendix A.

However, an exact solution to Eq. (7.10) is generally intractable. Therefore, approximate inference is employed for estimation of the marginal probability by applying the loopy belief propagation (LBP) algorithm [303].

7.2.3 Active Learning

In pool-based active learning, we suppose that during training we have access to a labeled dataset $\mathcal{L} = \{(\mathbf{x}_{\ell_i}, \mathbf{x}^*_{\ell_i}, y_i)\}_{i=1}^{N_\ell}$, with N_ℓ video sequences and an unlabeled dataset $\mathcal{U} = \{(\mathbf{x}_{u_i}, \mathbf{x}^*_{u_i})\}_{i=1}^{N_u}$, with N_u video sequences. We assume that pairs of original $\mathcal{X}$ and privileged information $\mathcal{X}^*$ are always available during training for both labeled and unlabeled datasets and only the corresponding label y_i may be missing. Our method is an incremental pool-based active learning approach, where at each iteration the most informative sample from $\mathcal{U}$ is selected. That is, the model selects samples that minimize the class label uncertainty. First, we learn the a-HCRF+ classifier on the labeled dataset. Then, we iteratively select an unlabeled sample pair $u = (\mathbf{x}_u, \mathbf{x}^*_u)$ and obtain the class posterior $p(y_u|u; \mathbf{w})$. In particular, we use two different strategies for selecting an unlabeled sample and ask for its label.

The first selection criterion is the entropy $\mathcal{H}(y_u|u; \mathbf{w})$, which measures how uncertain the classifier is about the class label y_u on the unlabeled sample u. Therefore, the most uncertain sample that maximizes the entropy is selected:

$$\hat{u} = \arg\max_{u \in \mathcal{U}} \left(-\sum_{y_u} p(y_u|u; \mathbf{w}) \log p(y_u|u; \mathbf{w}) \right) . \tag{7.11}$$

The second selection criterion corresponds to the ratio of class posteriors [328]. We estimate the class posterior for each unlabeled observation u and every class. Then, for these two classes that exhibit the largest posterior values $y_1 = \arg\max_{y_u} p(y_u|u; \mathbf{w})$ and $y_2 = \arg\max_{y_u \neq y_1} p(y_u|u; \mathbf{w})$, respectively, we select the unlabeled sample u that minimizes the ratio between the largest class posteriors:

$$\hat{u} = \arg\min_{u \in \mathcal{U}} \frac{p(y_1|u; \mathbf{w})}{p(y_2|u; \mathbf{w})} . \tag{7.12}$$

The ratio of class posteriors criterion allows to select an observation that lies closer to decision boundary of the learned classifier. Specifically, the main steps of proposed pool-based active learning methodology are summarized in Algorithm 5.

7.3 Experimental Results

The experiments were conducted in four challenging publicly available human activity recognition datasets. Three different types of privileged information were used: audio signal, human pose, and semantic attribute annotation and two active selection criteria were applied: entropy and ratio of class posteriors.

Algorithm 5 Pool-based active learning using a-HCRF+

procedure ACTIVEHCRFPLUS($\mathcal{L}, \mathcal{U}, \mathcal{X}, \mathcal{X}^*, \mathcal{Y}$)
  $\mathbf{w} \leftarrow \arg\min_{\mathbf{w}} (-L(\mathbf{w}))$ ▷ Train a-HCRF+ on $\mathcal{L}$.
  while $\mathcal{U} \neq \emptyset$ **do**
    Select an unlabeled observation $\hat{u}$ according to Eqs. (7.11) or (7.12) and query its label y.
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{(\hat{u}, y_u)\}$; ▷ Update labeled dataset $\mathcal{L}$.
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\hat{u}\}$; ▷ Update unlabeled dataset $\mathcal{U}$.
  end while
  $\mathbf{w} \leftarrow \arg\min_{\mathbf{w}} (-L(\mathbf{w}))$ ▷ Update a-HCRF+ parameters.
end procedure

7.3.1 Datasets

Parliament dataset [5]: This dataset contains 228 video sequences, depicting political speeches in the Greek parliament. The video sequences are labeled with one of three behavioral categories: *friendly*, *aggressive*, and *neutral*. The subjects express their opinion on a specific law proposal and they adjust their body movements and voice intensity level according to whether they agree with that or not.

TV human interaction (TVHI) dataset [6]: The TVHI dataset is a group of 300 video sequences collected by different TV shows and contain four kinds of interactions: *high fives*, *hugs* and *kisses*, which are equally distributed to the four classes (50 video sequences for each class). Negative examples (e.g., clips that do not contain any of the aforementioned interactions) consist the remaining 100 videos.

Two-person interaction (TPI) dataset [7]: This dataset consists of approximately 300 video sequences depicting two-person interactions captured by a Microsoft Kinect sensor. The sequences are categorized in eight different interaction classes including *approaching*, *departing*, *kicking*, *pushing*, *shaking hands*, *hugging*, *exchanging objects*, and *punching*. It also contains three-dimensional coordinates of 15 joints for each person at each frame.

Unstructured social activity attribute (USAA) [8]: The USAA dataset contains around 100 videos per class for training and testing, while it includes eight different semantic class videos of social occasions such as *birthday party*, *graduation party*, *music performance*, *non-music performance*, *parade*, *wedding ceremony*, *wedding dance*, and *wedding reception*. Each video is annotated with 69 attributes, which can be broken down into five broad classes: actions, objects, scenes, sounds, and camera movement.

7.3.2 Implementation Details

As video representation for all datasets, we used spatio-temporal interest points (STIP) [97]. Furthermore, for the Parliament and TVHI datasets, we extracted the mel-frequency cepstral coefficients (MFCC) [308] features along with their first and second order deriva-

tives. Audio features are also used as privileged information for these datasets. For the TPI dataset, we used the provided poses as privileged information, and for the USAA dataset we used the provided attribute annotation as privileged information. The number of hidden states was estimated based on cross validation, varying their from 3 to 20. The L_2 regularization scale term σ for was set to 10^k, with $k \in \{-3, \ldots, 3\}$. The proposed model was trained with a maximum of 400 iterations for the termination of the LBFGS optimization method.

For each dataset we used 5-fold cross validation to split into training and test sets. Finally, the initial training set was split into labeled and unlabeled set so that the size of the unlabeled set may vary from 10% to 50% of the total size of the original training set and the remaining videos form the labeled training set.

According to which selection criterion is employed (entropy or ratio of class posteriors), we proposed two variants of the method, called a-HCRF+ (entropy) and a-HCRF+ (ratioCP). We compared the proposed method with several baseline methods that may or may not use privileged information and/or active learning. First, we compared it with ordinary SVM [25] and HCRF [27], as if they could access both the original and the privileged information at test time. We also compared with state-of-the-art methods that employ privileged information such as SVM+ [313] (see Appendix B for more details), the rank transfer SVM+ (rt-SVM+) [317], which exploits a max-margin technique to transfer knowledge from the privileged to the original feature space, and the method of Wang and Ji [316], which exploits a loss inequality regularization (LIR) to address the sensitiveness of the loss function against the inequality constraints. However, these methods do not employ active learning, thus, we also compare with the method of Druck *et al.* [327], which applies generalized expectation criteria such as entropy (GEE) to select the most uncertain samples. Finally, we transformed standard SVM to an active learning based method (a-SVM) using entropy as selection criterion. For the SVM-based methods we consider a one-versus-one decomposition of multi-class classification scheme and average the results for every possible configuration, while the optimal parameters were selected using cross validation.

7.3.3 Results and Discussion

We assess the impact of privileged active learning by measuring the classification accuracy of both variants of the proposed method with varying number of unlabeled data. The obtained results are depicted in Figure 7.2. We may observe that for all datasets both pool-based active learning variants (entropy and ratio of class posteriors) always have superior performance than GEE and a-SVM methods as the size of unlabeled training observations increases. Specifically, for the TVHI dataset GEE may perform better only for the a-HCRF+ (ratioCP) variant, while for the USAA dataset a-HCRF+ (ratioCP) and a-SVM achieve similar results. This indicates the strength of the proposed privileged active learning method to recognize human actions from weakly labeled data without loosing accuracy due to the uncertainty of the model about class of each observation.

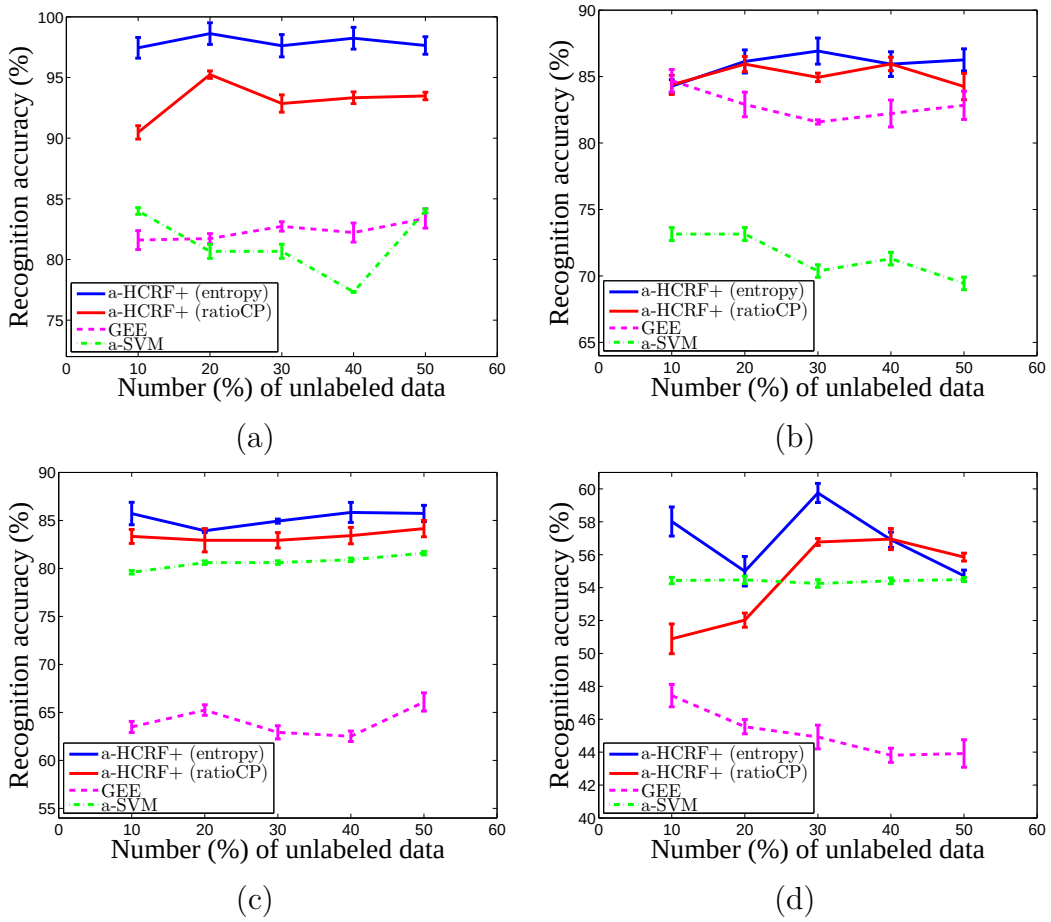


Figure 7.2: Comparison of classification accuracies with respect to the number of unlabeled data for (a) the Parliament [5], (b) the TVHI [6], (c) the TPI [7], and (d) the USAA [8] datasets.

Detailed results of the proposed method compared with state-of-the-art methods are presented in Table 7.1. We may observe that for all four datasets the proposed a-HCRF+ (entropy) method outperforms the state-of-the-art. For this variant, the classification performance significantly increased with respect to the LUPI-based SVM+ method for all datasets (e.g., 20% improvement of the Parliament dataset). Moreover, significant improvement is obtained, when the proposed method is compared to the active learning counterpart methods. Furthermore, the performance of the a-HCRF+ (ratioCP) variant achieves similar results to its counterpart that uses entropy as a selection criterion. Although the ratio of class posteriors for the Parliament and TVHI datasets may perform worse than standard HCRF model the overall performance is still better than the other methods. This is because of the presence of closely related classes as for some observation close to the decision boundary between two classes the logarithmic ratio of class posteriors may approach zero.

The corresponding confusion matrices for the a-HCRF+ (entropy) variant for the

Table 7.1: Comparison of the classification accuracies (%) for the Parliament [5], TVHI [6], TPI [7], and USAA [8] datasets. The results were averaged for all different configurations (mean ± standard deviation).

Method	Parliament [5]	TVHI [6]	TPI [7]	USAA [8]
Methods without privileged information and without active learning				
HCRF [27]	97.6 ± 0.6	81.3 ± 0.7	81.4 ± 0.8	54.0 ± 0.8
SVM [25]	72.6 ± 0.4	75.9 ± 0.6	79.4 ± 0.4	47.4 ± 0.1
Methods without privileged information and with active learning				
GEE [327]	82.3 ± 0.6	83.8 ± 0.8	66.1 ± 0.7	45.4 ± 0.6
a-SVM	80.5 ± 0.3	71.5 ± 0.5	80.6 ± 0.2	54.4 ± 0.2
Methods with privileged information and without active learning				
SVM+ [313]	78.4 ± 0.2	75.0 ± 0.2	79.4 ± 0.3	48.5 ± 0.1
rt-SVM+ [317]	57.7 ± 0.4	65.2 ± 0.1	56.3 ± 0.2	56.3 ± 0.2
LIR [316]	59.2 ± 0.2	74.8 ± 0.2	62.4 ± 0.3	48.5 ± 0.2
Methods with privileged information and with active learning				
a-HCRF+ (entropy)	$\mathbf{98.1} \pm 0.9$	$\mathbf{85.8} \pm 0.5$	$\mathbf{85.2} \pm 0.6$	$\mathbf{56.9} \pm 0.4$
a-HCRF+ (ratioCP)	93.0 ± 0.2	85.1 ± 0.8	83.8 ± 1.0	55.2 ± 0.5

best split for each dataset are shown in Figure 7.3. It is worth mentioning that for the Parliament and TVHI datasets the classification errors between different classes are relatively small. For the TPI dataset, only a few classes are highly correlated to each other (e.g., the class *shake hands* is confused with the classes *push* and *hug*). On the other hand, the USAA dataset, shows high confusion between the different classes (e.g., *wedding ceremony* is confused with the class *birthday party*). This is because of the large intra-class variabilities, since different classes may have similar attribute representation of human actions.

7.4 Conclusion

In this chapter, the problem of human activity recognition in a semi-supervised framework is investigated. A combination of learning using privileged information and active learning into a unified framework indicated that human actions can effectively be recognized. Moreover, two variants of the proposed a-HCRf+ method were proposed. The fist uses entropy as a measure of uncertainty of the actual class of unlabeled observations and the second selects an unlabeled observation that lies closer to the decision boundary. Several types of auxiliary information were used indicating that the proposed method is not limited to a specific form of privileged information. The experimental results on four different publicly available datasets were very promising and supported the fact that both

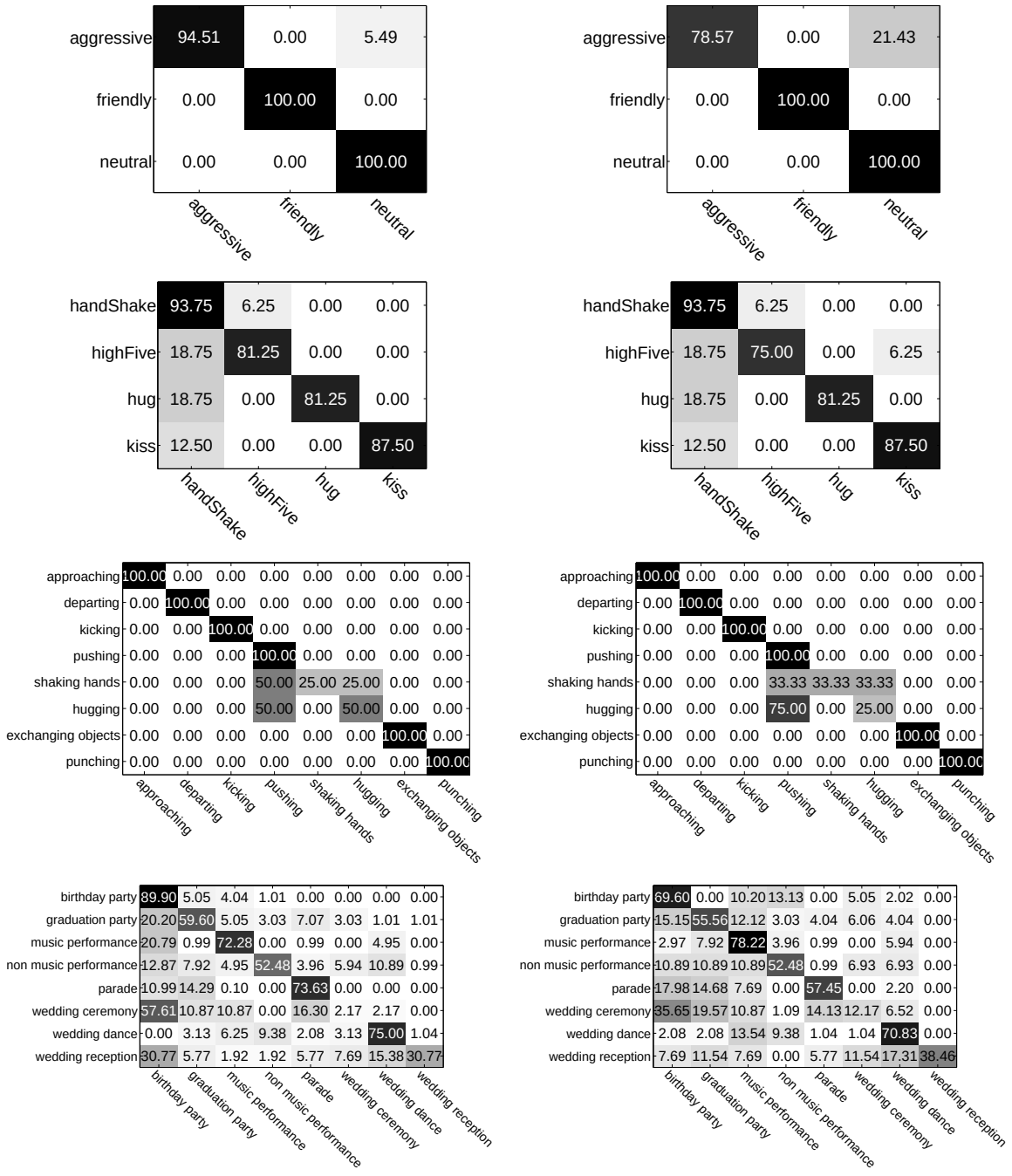


Figure 7.3: Confusion matrices for the classification results for the best split of the proposed a-HCRF+ model for the Parliament [5] (first row), the TVHI [6] (second row), the TPI [7] (third row), and the USAA [8] (fourth row) datasets. Right column corresponds to a-HCRF+ (entropy) and left column corresponds to a-HCRF+ (ratioCP) variants, respectively.

LUPI and active learning schemes, when used together, achieve superior performance than the state-of-the-art.

Chapter 8

Exploiting Privileged Information for Facial Expression Recognition

8.1 Introduction

8.2 Leaning to Transfer Privileged Information

8.3 Experimental Results

8.4 Conclusion

8.1 Introduction

Facial expression recognition has recently attracted much attention due to its applicability in several fields of biometrics, computer vision, and machine learning [42, 333]. Its applications may vary from video surveillance, driver and/or patient monitoring to human-machine interactions. Many facial expression recognition systems provide information about the personality and psychological state of a person. In real world, humans express their emotions as a combination of verbal and non-verbal multimodal cues such as gestures, facial expressions and auditory cues. Combining different modalities poses a great challenge on recognizing facial expressions [121, 334].

The multimodal nature of the problem requires the development of new learning techniques. Several approaches such as multi-task learning [335] and domain adaptation [336] have been proposed for dealing with multimodal problems. In multi-task learning the goal is to improve the performance across all tasks, while domain adaptation methods consider individual domains, which are combined to improve the performance on a target domain. These approaches assume that the classifier is trained and tested on similar sets of data. However, exploiting the same type of information during training and testing may not always be possible due to data acquisition constraints. To this end, learning

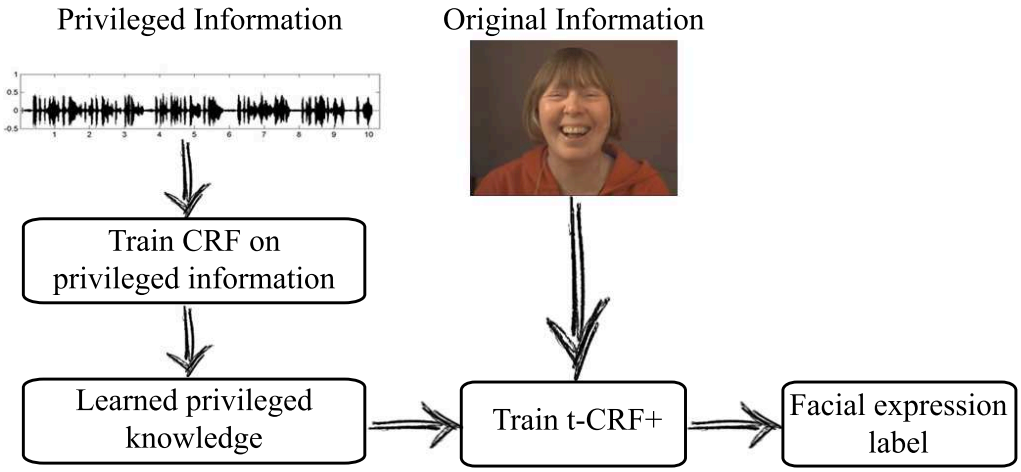


Figure 8.1: An overview of the proposed framework.

using privileged information [313] has been explored to cope with the inhomogeneity in training and testing information. The idea of privileged information is that one may have access to additional information about the training samples, which is not available during testing. However, defining which information may be considered as privileged and which as regular is not an easy task as the problem is not straightforward [337], while the lack of informative data or the presence of misleading information may influence the performance of the model by introducing bias.

In this chapter, we address these limitations by introducing a novel probabilistic model, which incorporates the LUPI paradigm into a unified framework for recognizing facial expressions and affective states of a person. We propose an efficient method to indirectly transfer the knowledged from privileged to the original feature space using conditional random fields (CRFs) [26], called transfer-CRF+ (t-CRF+). Specifically, the privileged information is provided as additional input to our model through a two step classification process. We first train a standard CRF model on the privileged data and encode the ability of privileged information to distinguish between different class labels into the model weights. The learned privileged weights are then used to penalize the training process on the original feature space by learning the conditional probability distribution between the class labels and original observations. The penalty term encourages the model to assign larger weights to samples that have a good evidence to distinguish between classes both in privileged and original feature space and smaller weights to the contrary. In other words, the proposed model is able to enhance the classification accuracy by learning a better estimate of model parameters in the original feature space by transferring the knowledge from the privileged data. Figure 8.1 illustrates an overview of the proposed methodology.

The main contributions of this work can be summarized in the following points: (i) a new probabilistic classification scheme based on CRFs is proposed to improve the recognition of facial expressions and affective states of a person by gaining additional knowledge about the training data using privileged information; (ii) information transferring is used to keep only the relevant information between privileged and original feature space. Note

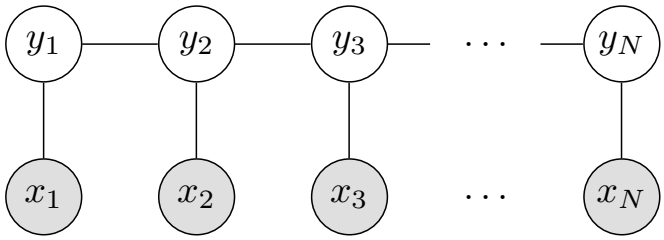


Figure 8.2: Graphical representation of the chain structure CRF model. The grey nodes are the observed features (x_i) and the white nodes are unknown labels (y_i), respectively.

that the proposed method is general and is not limited to the use of any specific form of privileged information, but rather it is general for any form of additional data.

8.2 Leaning to Transfer Privileged Information

We consider a labeled dataset with N video sequences, which instead of paired input-output samples, it consists of triplets $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^{M_\mathbf{x}}$ is a training observation from the feature space $\mathcal{X}$ and y_i corresponds to a class label defined in a finite label set $\mathcal{Y}$. In the context of learning using a privileged information paradigm, additional information about the observations $\mathbf{x}_i$ is encoded in a feature vector $\mathbf{x}_i^* \in \mathbb{R}^{M_{\mathbf{x}^*}}$ in the privileged space $\mathcal{X}^*$. Such privileged information is provided only at the training step and it is not available during testing, while no further assumption about the form of the privileged data is made.

In particular, $\mathbf{x}_i^*$ does not necessarily share the same characteristics with the original data, but is rather computed as a very different kind of information, which may contain verbal and/or non-verbal multimodal cues such as (i) visual features, (ii) attributes, (iii) textual descriptions of the observations, (iv) image/video tags, and (vi) audio cues. The goal of LUPI is to use the privileged information $\mathbf{x}_i^*$ as a medium to construct a superior classifier for solving practical problems than one would learn without it.

8.2.1 t-CRF+ Model Formulation

The proposed method uses CRFs, which are defined by a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 8.2), as the probabilistic framework for modeling the facial expressions of a subject in a single image or video. During training, a classifier and the mapping from observations to the label set for the different configurations are learned. In testing, a probe sequence is classified into its respective state using belief propagation (BP) [303].

The CRF model is a member of the exponential family and the probability of the class label given an observation sequence is given by:

$$p(y|\mathbf{x}; \mathbf{w}) = \exp\left(E(y|\mathbf{x}; \mathbf{w}) - A(\mathbf{w})\right) , \tag{8.1}$$

where $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters. We assume that our model follows the first-order Markov chain structure (i.e., the current state affects the next state). Finally, $E(y|\mathbf{x}; \mathbf{w})$ is a function of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

$$A(\mathbf{w}) = log \sum_{y'} \exp\left(E(y'|\mathbf{x}; \mathbf{w})\right) . \tag{8.2}$$

Different sufficient statistics $E(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ in (8.1) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$E(y|\mathbf{x}; \mathbf{w}) = \sum_{j \in \mathcal{V}} \Phi(y_j, \mathbf{x}_j; \boldsymbol{\theta}) + \sum_{j,k \in \mathcal{E}} \Psi(y_j, y_k; \boldsymbol{\omega}) , \tag{8.3}$$

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned.

The unary potential is expressed by:

$$\Phi(y_j, \mathbf{x}_j; \boldsymbol{\theta}) = \sum_{j} \sum_{a \in \mathcal{Y}} \boldsymbol{\theta}^\top \mathbb{1}(y_j = a)\mathbf{x}_j , \tag{8.4}$$

and it can be seen as an observation feature function, which models the relationship between the label y_j and the observations $\mathbf{x}_j$, where $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1, if its argument is true and 0 otherwise.

The pairwise potential is a transition function and represents the association between a pair of connected labels y_j and y_k. It is expressed by:

$$\Psi_\ell(y_j, y_k; \boldsymbol{\omega}_\ell) = \sum_{a,b \in \mathcal{Y}} \sum_{\ell} \boldsymbol{\omega}_\ell \mathbb{1}(y_j = a)\mathbb{1}(y_k = b) . \tag{8.5}$$

Note that the CRF model keeps a transition matrix for each label.

8.2.2 Parameter Learning and Inference

In the classical CRF model, the optimal parameters are estimated during training by maximizing the following loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w}) - \frac{\|\mathbf{w}\|^2}{2\sigma^2} . \tag{8.6}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}; \mathbf{w})$ and quantifies how well the distribution in Eq. (8.1) defined by the parameter vector $\mathbf{w}$ matches the labels y. The second term is a Gaussian prior with variance σ^2 and works as a regularizer.

Our work is based on the intuition that privileged information is more informative than the ordinary information and thus, learning on privileged data may improve the classification. The proposed t-CRF+ model relies on the idea that instead of jointly learning the ordinary and privileged information, we first train an ordinary CRF on the

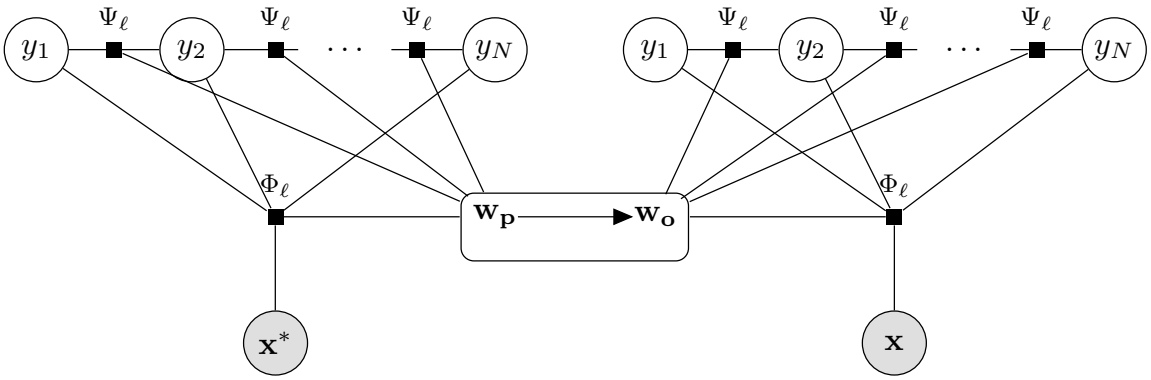


Figure 8.3: Proposed t-CRF+ model. First, a standard chain structure CRF model is trained on the privileged feature space ($\mathcal{X}^*$) with parameters $\mathbf{w_p}$. Then, the privileged knowledge is transferred to the original feature space ($\mathcal{X}$). The square nodes correspond to the unary and pairwise potentials, which are conditioned on their hyper-parameters $\mathbf{w_p}$ and $\mathbf{w_o}$, respectively.

privileged feature space $\mathcal{X}^*$, and then we exploit the obtained knowledge to improve the performance on the target feature space $\mathcal{X}$, for which training data are always available during training and testing.

To achieve the knowledge transfer, we penalize the loss function of the standard CRF model with an additional term that corresponds to a Gaussian prior with zero mean and variance σ_p^2. Thus, the loss function in Eq. (8.6) is modified to encode the knowledge transfer from privileged to original feature space:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w_o}) - \frac{\|\mathbf{w_p} - \mathbf{w_o}\|^2}{2\sigma_p^2} - \frac{\|\mathbf{w_o}\|^2}{2\sigma_o^2}\,, \tag{8.7}$$

where $\mathbf{w_o}$ and $\mathbf{w_p}$ are the model parameters when training in the original and the privileged feature space, respectively. In Eq. (8.7), the parameters $\mathbf{w_o}$ and $\mathbf{w_p}$ should be of equal length and this is achieved using canonical correlation analysis (CCA) [228] as a preprocessing step. The parameters σ_p^2 and σ_o^2 are tuning parameters that control the degree of influence of the privileged and the original information, respectively.

Figure 8.3 illustrates the graphical representation of the proposed t-CRF+ model. The t-CRF+ model is parameterized by two hyper-parameters $\mathbf{w_p}$ and $\mathbf{w_o}$. In this case, the privileged information is indirectly transferred for learning the baseline CRF model through the learned prediction function for each training instance in the privileged space. The privileged parameters $\mathbf{w_p}$ are used in the original conditional log-likelihood function to influence the values of the parameters in the original feature space.

The degree of influence the privileged information may have upon the original information depends on the degree of evidence for each privileged weight. The smallest the values of the privileged weights $\mathbf{w_p}$ are, the smallest the influence of privilege data also is. The opposite occurs when samples with larger privileged weights $\mathbf{w_p}$ may contribute more heavily through the Gaussian prior in Eq. (8.7) and thus, the privileged knowledge may have greater effect on the finally parameter learning. This process can be viewed as

Algorithm 6 Transferring knowledge from $\mathcal{X}^*$ to $\mathcal{X}$ using t-CRF+

Input: Original data $\mathcal{X}$, privileged data $\mathcal{X}^*$, class labels $\mathcal{Y}$.

Output: Predicted labels.

1: Perform canonical correlation to make the dimensions of $\mathcal{X}$ and $\mathcal{X}^*$ equal.
2: Train a standard CRF on the privileged data $(\mathbf{x}^*, y)$ using Eq. (8.6) and estimate models' parameters $\mathbf{w_p}$.
3: Train a CRF on the original feature space $(\mathbf{x}, y)$ using Eq. (8.7) to transfer the knowledge from the privileged to the original feature space.
4: Obtain final labels using Eq. (8.8).

selection process, where the most informative data in the privileged space contribute to the classification of the true label.

In our implementation, the loss function in Eq. (8.7) is optimized using a gradient-descent optimization method. More specifically, we used the limited-memory BFGS (LBFGS) method [306] to minimize the negative log-likelihood of the data.

Having computed the optimal parameters $\mathbf{w}^*$ in the training step, our goal is to estimate the optimal label configuration over the testing input, where the optimality is expressed in terms of a cost function. To this end, we maximize the posterior probability:

$$y = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}; \mathbf{w}) \,. \tag{8.8}$$

The marginal probability is obtained by applying the BP algorithm [25] using the graphical model as depicted in Fig. 8.2. The main steps of the proposed t-CRF+ classification model are summarized in Algorithm 6.

8.3 Experimental Results

To show the ability of the proposed t-CRF+ method to generalize, we compared it with several state-of-the-art methods for two different computer vision applications, namely emotional facial recognition, and facial expression recognition, with different type of privileged information for each problem. For the first problem, we used the AVEC 2011 dataset [9] and for the second we used the extended Cohn-Kanade (CK+) dataset [10].

8.3.1 Datasets

AVEC 2011 audio/visual challenge dataset [9]: This dataset consists of 95 sequences of upper body video segments at resolution of 780×580 at 49.979 fps while the audio was recorded at 48 kHz, and is part of the SEMAINE corpus [338]. The AVEC 2011 dataset consists of 31 videos for training, 32 videos for validation, and 32 videos for testing, annotated with four affective labels such as activation, expectation, power, and valence. As original features, we used the pre-computed video features provided by the

dataset, and the privileged information was selected to be the provided audio features, which were obtained from various low-level descriptors. Due to the large amount of data and relatively high feature dimensionality for this dataset, we followed the same strategy as proposed by Schuller *et al.* [9] for sub-sampling the data and reducing the feature dimension.

Cohn-Kanade (CK+) dataset [10]: This dataset describes facial expressions such as anger, disgust, fear, happiness, sadness, surprise, and contempt. All facial expressions are expressed by the facial action coding system (FACS) [230], which describes all possible facial expressions as a combination of action units (AU), extracted from each participant, to identify their emotional state. It consists of 593 video sequences of 123 subjects captured from the neutral face to the peak expression. Since FACS are coded only at the peak frame, we only considered the peak frame in our experiments. For this dataset, the original features were selected to be the 68 tracked facial landmarks obtained by active appearance models [339] and the privileged information was selected to be the 17 annotated action units, all provided by the database creators.

8.3.2 Baseline Approaches

We compared the proposed method with several baseline methods that may or may not use privileged information. First, we used SVM+ [313], which consists of optimizing the hyperplane parameters such that it can minimize the probability of incorrect classifications and increase the convergence rate. A brief description of SVM+ can be found in Appendix B. The second baseline is the rank transfer SVM+ (rt-SVM+) [317], which exploits a max-margin technique to transfer knowledge from the privileged to the original feature space. Finally we compared with the method of Wang and Ji [316], which exploits a loss inequality regularization (LIR) to address the sensitiveness of the loss function against the inequality constraints.

We also compared the proposed t-CRF+ method with ordinary SVM and CRF, as if they could access both the original and the privileged information at test time. This means that we do not differentiate between regular and privileged information, but use both forms of information as regular to infer the underlying class label instead. In this case, we considered early fusion to combine features from different modalities. Furthermore, to complete the study, we also trained an CRF model that uses only the regular and only the privileged information for training and testing.

8.3.3 Model Selection

The L_2 regularization scale terms σ_p and σ_o were set to 10^k, with $k \in \{-3, \ldots, 3\}$. The optimal parameters for all baseline methods were selected using cross validation, and the best parameters or parameter sets were used to retrain the model. Finally, our model in Eq. (8.7) was trained with a maximum of 400 iterations for the termination of the LBFGS minimization method.

Dataset	Regular	Privileged	Accuracy (%)	AUC (%)
AVEC 2011 [9]	visual	✗	60.5	85.7
	audio	✗	59.6	83.1
	visual+audio	✗	60.7	70.6
	visual	audio	**70.7**	**91.2**
CK+ [10]	facial lnd	✗	85.4	91.9
	AU	✗	85.1	92.5
	facial lnd + AU	✗	85.9	93.4
	facial lnd	AU	**93.6**	**99.3**

Table 8.1: Comparison of feature combinations for classifying facial expressions and affective states on AVEC 2011 [9], and CK+ [10] datasets. The crossmark indicates the absence of privileged information during training.

The evaluation of our method was performed using 5-fold cross validation to split the datasets into training and test sets, according to the documentation described in each dataset, and we report the average results over all the examined configurations. For the SVM-based methods we consider a one-versus-all decomposition of multi-class classification scheme and average the results for every possible configuration.

8.3.4 Results and Discussion

In the fist set of experiments, we assessed the impact of privileged information to recognize affective states of emotional audio and video dyadic interactions between human participants using the AVEC 2011 dataset [9], and we also trained the proposed model to the CK+ dataset [10] for recognizing facial expressions. For the evaluation of the proposed method we used the classification accuracies and the area under the ROC curve (AUC), which compares the true positive against the false positive rate. The benefit of using robust privileged information along with conventional data instead of using each modality separately or both modalities as regular information is shown Table 8.1. For the classification, we used a standard CRF model and compared it with the proposed t-CRF+ method. We may observe that for both datasets, if only privileged information is used as regular features for classification both the classification accuracy and the AUC are lower than when using only the regular information for the classification task. However, these results are relatively similar to each other, which leads to the conclusion that finding proper privileged information is not always a straightforward procedure. Moreover, the proposed classification scheme performs better than all other approaches. These results demonstrate that the t-CRF+ model can successfully exploit the privileged information to improve the recognition accuracy.

In the second set of experiments, the proposed approach was compared with several state-of-the-art methods, that may or may not use privileged information for both

Method	AVEC 2011		CK+	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Methods without privileged information				
SVM [25]	57.3 ± 0.111	73.7 ± 0.287	84.8 ± 0.079	87.3 ± 0.095
CRF [26]	60.7 ± 0.825	70.6 ± 0.408	85.9 ± 0.626	93.4 ± 0.095
Methods with privileged information				
rt-SVM+ [317]	63.6 ± 0.069	86.3 ± 0.138	85.7 ± 0.103	88.4 ± 0.147
SVM+ [313]	59.6 ± 0.041	65.7 ± 0.134	87.7 ± 0.083	85.6 ± 0.080
LIR [316]	49.3 ± 0.066	67.2 ± 0.196	87.3 ± 0.834	85.5 ± 0.081
t-CRF+	$\mathbf{70.7} \pm 0.273$	$\mathbf{92.9} \pm 0.024$	$\mathbf{93.6} \pm 0.667$	$\mathbf{99.3} \pm 0.008$

Table 8.2: Comparison of the classification accuracies and the area under the ROC curve (%) for the AVEC 2011 [9] and the CK+ [10] datasets.

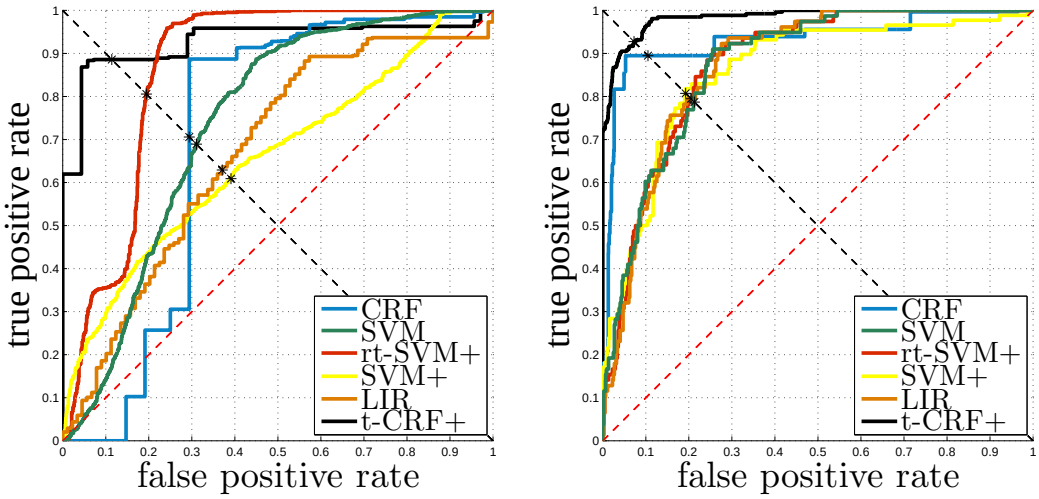
datasets. The results are presented in Table 8.2. The results indicate that our approach improved the classification accuracy and the AUC. On AVEC 2011, we significantly managed to increase the classification accuracy by approximately 10% and the AUC by 20% with respect to CRF and SVM, which do not employ privileged information, as our approach achieves very high recognition accuracy for this dataset (70.7%). The improvement of our method compared to the methods that also employ privileged information is high. Furthermore, our method outperforms by approximately 7% in recognition accuracy and by 5% in AU the rt-SVM+, which also employs transferring of privileged information. Accordingly, for the CK+ dataset, the improvement against the state-of-the-art methods is also high and almost 8% higher accuracy with respect to the achieved by rt-SVM+ and 6% higher when compared to SVM+ and LIR methods. We may also observe that for this dataset, the AUC values achieved by the proposed t-CRF+ model are very high and close to the ideal classifier. In general, the significantly high increase in all evaluation indices by our model indicates the strength of the proposed method.

In order to provide a statistical evidence of the recognition results, we computed the p-values of the obtained results with respect to the compared methods. The null hypothesis was defined as: the mean performances (accuracies or AUC) of the proposed model are equal to the state-of-the-art methods; and the alternative hypothesis was defined as: the mean performances (accuracies or AUC) of the proposed model are higher than those of the state-of-the-art methods. For the assessment of the statistical significance, we used paired t-tests with statistical significance threshold $p < 0.05$ for all experiments. The resulted p-values for both datasets are reported in Table 8.3. According to these results, we conclude that for both datasets the the null hypothesis is rejected as the p-values were less than the significance level of 0.05, and thus, the improvements obtained by our model are statistically significant and not due to chance.

The corresponding ROC curves for both datasets are depicted in Fig. 8.4. The red

Method	AVEC 2011		CK+	
	Accuracy	AUC	Accuracy	AUC
SVM [25]	0.0174	0.0383	0.0257	0.0089
CRF [26]	0.0435	0.0145	0.0390	0.0851
rt-SVM+ [317]	0.0269	0.7683	0.0361	0.0001
SVM+ [313]	0.0035	0.0062	0.0776	0.0026
LIR [316]	0.0043	0.0054	0.0666	0.0025

Table 8.3: p-values of the proposed method for the AVEC 2011 [9] and the CK+ [10] datasets.



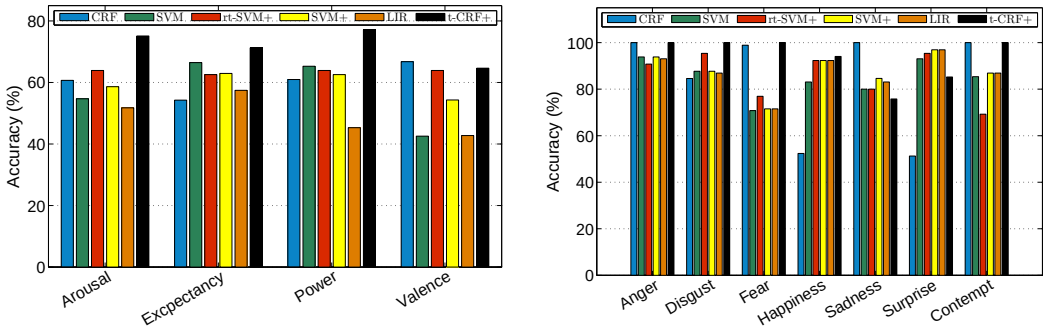
(a) (b)

Figure 8.4: Illustration of ROC curves for (a) AVEC 2011 [9] and (b) CK+ [10] datasets.

dotted diagonal line corresponds to complete random guess. The intersection of the ROC curve for each method with the black diagonal line, corresponds to the equal error rate (EER). We may see that for the AVEC 2011 dataset the proposed method has the lowest EER (0.1141) and for the CK+ the EER is 0.0726, which is smaller than the state-of-the-art methods.

Finally, the classification performance of the proposed method against the baseline methods for each class separately on both datasets is depicted in Fig. 8.5. We may observe that for AVEC 2011 in three out of four classes the proposed t-CRF+ method has the highest accuracy. However, for the valence class the standard CRF model performs slightly better, but still our method outperforms the rest of the state-of-the-art. For the CK+ dataset, the classification accuracy on four classes is perfect (100%), but for the classes sadness and surprise the proposed method performs worse than the baseline methods, mostly because some action units are hard to detect.

In general, our method is able to transfer privileged information to the original space



(a) (b)

Figure 8.5: Comparison of recognition performance accuracies (%) of each class for (a) AVEC 2011 [9] and (b) CK+ [10] datasets.

in a more efficient way than SVM+, rt-SVM+, and LIR. We can also observe that the proposed method outperforms both the SVM and CRF models. However, the information that is being transferred may not always improve the classification in all classes, although the classification results in each class are relatively high, as it is mainly a matter of training and testing set size and the quality/structure of the data.

8.4 Conclusion

In this chapter, the problem of facial expression recognition in the framework of learning using privileged information is addressed. It is demonstrated that the proposed t-CRF+ method is able to efficiently exploit additional information about the training data to transfer the knowledge learned from privileged to the original feature space for predicting the true class. In contrast to conventional classification tasks, it is observed that the use of privileged information can lead to superior performance in classifying facial emotions for both accuracy and AUC indices. Moreover, various forms of data that can be used as privileged were investigated. Experimental results on different publicly available benchmarks showed improvements over state-of-the-art methods that may or may not employ privileged information.

Chapter 9

Conclusions and Future Work

9.1 Conclusions

9.2 Limitations and Future Work

9.1 Conclusions

In this thesis, we carried out a comprehensive study of state-of-the-art methods of human activity recognition and proposed a hierarchical taxonomy for classifying these methods. We surveyed different approaches, which were classified into two broad categories according to the source channel each of these approaches employ to recognize human activities. We discussed unimodal approaches and provided an internal categorization of these methods, which were developed for analyzing gesture, atomic actions, and more complex activities, either directly or employing activity decomposition into simpler actions. We also presented multimodal approaches for the analysis of human social behaviors and interactions. We discussed the different levels of representation of feature modalities and reported the limitations and advantages for each representation. A comprehensive review of existing human activity classification benchmarks was also presented and we provided the characteristics of building an ideal human activity recognition system.

Based on the above observations, our work in Chapter 3 was focused on recognizing atomic actions and more complex human activities such as sport activities by tracking optical flow features in time. The obtained trajectories were grouped to represent an individual class of human action with a compact set of motion curves. Although this approach may perform relatively well for datasets with small variations in the backgrounds, it may not achieve equally high results for more complex and dynamic scenes mainly due to the erroneous nature of optical flows. A possible extension to this problem would be the use of a "hyper-cluster" that may capture the outliers occurred form data acquisition. Moreover, instead of clustering trajectories of similar action classes from the whole body,

an interesting approach would be to cluster motion trajectories of similar body parts (e.g., left/right hand, torso, and head) separately.

In Chapter 4, a new challenging dataset (*Parliament*) was introduced for recognizing high-level human activities and behaviors. First, we presented a method for recognizing human behaviors that used a fully connected CRF model, where different labels for each video frame were considered. Although this representation makes the model more suitable to handle video sequences with more than one label per video, it significantly increases the complexity of the model. To this end, in Chapter 5, we replaced the above model with a chain-structured HCRF model and employed a feature selection technique along with voice features to improve recognition accuracy.

Standard human activity classification systems assume that both training and testing sequences represent similar types of information. However, in real-world applications, this may not always be possible due to data acquisition constraints. In Chapter 6, we developed a solution to this limitation by presenting an improved version of the HCRF model that incorporates the LUPI paradigm and it is able to handle auxiliary (privileged) information about the input data, which is accessible only during training but never during testing. The proposed method is not tightly combined to a specific type of privileged information, but it can cope with different types of auxiliary data. Both maximum likelihood and max-margin approaches were used to train the proposed model, while the regularization parameters for both approaches were iteratively estimated through a self-training procedure. The results indicated that high recognition rates were achieved and we also managed to beat the state-of-the-art in the LUPI framework. However, in the current work, the number of hidden states is determined using cross validation. Learning the number of the hidden variables necessitates more complex models and is a topic that needs of further exploration.

An extension of the aforementioned method that incorporates both LUPI framework and active learning to take advantage of semi-supervised learning was also proposed in Chapter 7. Training data were considered to be both labeled and unlabeled, while in testing data privileged information is not available. Entropy and the distance form the decision boundary were used to selected the most informative unlabeled sample and obtain its label. Although both selection criteria were individually found to work relatively well, a rough combination of them may not achieve equally close results as different weights to each criterion should be assigned. Thus, a possible extension of this work should be the investigation of other query selection criteria and how active learning can be used to recognize actions from unsegmented sequences.

We also investigated how privileged information could be applied not only to human activity recognition but in other applications in biometrics such as facial expression recognition. In Chapter 8, privileged information was embedded into a chain-structured CRF model to transfer privileged knowledge to the original feature space by penalizing the models' weights using a Gaussian prior over the privileged space. In the current work, the proposed model considers that privileged information consists of one modality. Thus, the

evaluation of our method on multiple and heterogeneous sources of privileged information at the same time is an issue that needs to be investigated.

The gap of a complete representation in number of human activities and the corresponding data collection and annotation is still a challenging and unbridged problem. In particular, we may conclude that despite the tremendous increase of human understanding methods, many problems still remain open, including modeling of human poses, handling occlusions, and annotating data.

9.2 Limitations and Directions for Future Work

Besides the vast amount of research in the field of activity recognition, a generalization of the learning framework is crucial towards modeling and understanding real world human activities. Several challenges that correspond to the ability of a classification system to generalize under external factors, such as variations in human poses and different data acquisition, are still open issues. The ability of a human activity classification system to imitate humans' skill in recognizing human actions in real time is a future challenge to be tackled. Machine learning techniques that incorporate knowledge-driven approaches may be vital for human activity modeling and recognition in unconstrained environments, where data may not be adequate or may suffer from occlusions and changes in illuminations and view point.

Training and validation methods still suffer from limitations such as slow learning rate, which gets even worse for large scale training data, and low recognition rate. Although much research focuses on leveraging human activity recognition from big data, this problem is still in its infancy. The exact opposite problem (i.e., learning human activities from very little training data or missing data) is also very challenging. Several issues concerning the minimum number of learning examples for modeling the dynamics of each class or safely inferring the performed activity label are still open and need further investigation. More attention should also be put in developing robust methods under the uncertainty of missing data either on training steps or testing steps.

The role of appropriate feature extraction for human activity recognition is a problem that needs to be tackled in future research. The extraction of low-level features that are focused on representing human motion is a very challenging task. To this end, a fundamental question arises: are there features that are invariant to scale and viewpoint changes, which can model human motion in a unique manner, for all possible configurations of human pose?

Furthermore, there exists a great need for efficiently manipulating training data that may come from heterogeneous sources. The number and type of different modalities that can be used for analyzing human activities is an important question. The combination of multimodal features such as body motion features, facial expressions, or the intensity level of voice may produce superior results, when compared to unimodal approaches, On the other hand, such a combination may constitute over-complete examples that can be

confusing and misleading. The proposed multimodal feature fusion techniques does not incorporate the special characteristics of each modality and the level of abstraction for fusing. Therefore, a comprehensive evaluation of feature fusion methods that retain the feature coupling is an issue that needs to be assessed.

The lack of large and realistic human activity recognition datasets is a significant challenge that needs to be addressed. An ideal action dataset should cover several topics, including diversity in human poses for the same action, a wide range of ground truth labels, and variations in image capturing and quality. Although a list of action datasets that correspond to most of these specifications has been introduced in the literature, the question of how many actions we can actually learn is a task for further exploration. Although most of the existing datasets contain no more than two tens of classes, there exist a few datasets having a few hundreds of classes. In such large datasets, the ability to distinguish between easy and difficult examples for representing the different classes and recognizing the underlying activity is difficult. This fact opens a promising research area that should be further studied.

Another challenge worthy of further exploration is the exploitation of unsegmented sequences, where one activity may succeed another. Frequent changes in human motion and actions performed by groups of interacting persons makes the problem amply challenging. More sophisticated high-level activity recognition methods need to be developed, which should be able to localize and recognize simultaneously occurring actions by different persons.

Appendix A

Conditional Distribution of the Privileged Information

A.1 Conditional Student's t-Distribution

A.1 Conditional Student's *t*-Distribution

Recall that $\mathbf{x} \in \mathbb{R}^{M_\mathbf{x} \times T}$ is an observation sequence of length T and $\mathbf{x}^* \in \mathbb{R}^{M_{\mathbf{x}^*} \times T}$ corresponds to the privileged information of the same length. We partition the original set $(\mathbf{x}^*, \mathbf{x})^T \in \mathbb{R}^{M \times T}$ into two disjoint subsets, where $\mathbf{x}^*$ forms the first $M_{\mathbf{x}^*}$ components of $(\mathbf{x}^*, \mathbf{x})^T \in \mathbb{R}^{M \times T}$ and $\mathbf{x}$ comprises the remaining $M - M_{\mathbf{x}^*}$ components. If the joint distribution $p(\mathbf{x}, \mathbf{x}^*)$ follows a Student's t-law, with mean vector $\mu = (\mu_{\mathbf{x}^*}, \mu_\mathbf{x})^T$, a real, positive definite, and symmetric $M \times M$ covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{x}^*\mathbf{x}^*} & \Sigma_{\mathbf{x}^*\mathbf{x}} \\ \Sigma_{\mathbf{x}\mathbf{x}^*} & \Sigma_{\mathbf{x}\mathbf{x}} \end{pmatrix}$ and $\nu \in [0, \infty)$ corresponds to the degrees of freedom of the distribution [321], then the conditional distribution $p(\mathbf{x}|\mathbf{x}^*)$ is also a Student's t-distribution:

$$\begin{aligned} p(\mathbf{x}^*|\mathbf{x}) &= \mathrm{St}(\mathbf{x}^*; \mu^*, \Sigma^*, \nu^*) \\ &= \frac{\Gamma\left(\left(\nu^* + M\right)/2\right) |\Sigma_{\mathbf{x}\mathbf{x}}|^{1/2}}{\left(\pi\nu^*\right)^{M_\mathbf{x}/2} \Gamma\left(\left(\nu^* + M_\mathbf{x}\right)/2\right) |\Sigma^*|^{1/2}} \cdot \frac{\left[1 + \frac{1}{\nu^*} \mathbf{x}^T \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \mathbf{x}\right]^{\frac{(\nu^* + M_\mathbf{x})}{2}}}{\left[1 + \frac{1}{\nu^*} Z^T \Sigma^{*-1} Z\right]^{\frac{(\nu^* + M)}{2}}} \,. \end{aligned} \tag{A.1}$$

The mean μ^*, the covariance matrix Σ^* and the degrees of freedom ν^* of the conditional distribution $p(\mathbf{x}^*|\mathbf{x})$, respectively, are computed by the respective parts of μ and Σ:

$$\mu^* = \mu_{\mathbf{x}^*} - \Sigma_{\mathbf{x}^*\mathbf{x}} \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \left(\mathbf{x} - \mu_\mathbf{x}\right) \,, \tag{A.2}$$

$$\Sigma^* = \frac{\nu_{\mathbf{x}^*} + \left(\mathbf{x} - \mu_\mathbf{x}\right)^T \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \left(\mathbf{x} - \mu_\mathbf{x}\right)}{\nu_{\mathbf{x}^*} + M_{\mathbf{x}^*}} \cdot \left(\Sigma_{\mathbf{x}^*\mathbf{x}^*} - \Sigma_{\mathbf{x}^*\mathbf{x}} \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \Sigma_{\mathbf{x}\mathbf{x}^*}\right) \,, \tag{A.3}$$

$$\nu^* = \nu_{\mathbf{x}^*} + M_{\mathbf{x}^*} \,. \tag{A.4}$$

The parameters (μ, Σ, ν) of the joint Student's t-distribution $p(\mathbf{x}^*, \mathbf{x})$, which are defined by the corresponding partition of the vector $(\mathbf{x}^*, \mathbf{x})^T$, are estimated using the expectation-maximization (EM) algorithm [321]. Then, the parameters of the conditional distribution $p(\mathbf{x}^*|\mathbf{x})$ are computed using Eq. (A.2)-(A.4).

It is worth noting that by letting the degrees of freedom ν^* to go to infinity, we can recover the Gaussian distribution with the same parameters. If the data contain outliers, the degrees of freedom parameter ν^* are weak and the mean and covariance of the data are appropriately weighted in order not to take into account the outliers.

Note that μ^* is a linear function of the observations $\mathbf{x}$ and it is the same as the conditional mean in the case that the sample data $\mathbf{x}^*$ and $\mathbf{x}$ follow a Gaussian distribution and Σ^* is influenced by the realization $\mathbf{x}$. If ν^* tends to reach infinity, we can approximate the Gaussian conditional covariance as the Student's t-distribution is a heavy tailed approximation to the Gaussian:

$$\lim_{\nu^* \to \infty} \Sigma^* = \Sigma_{\mathbf{x}^*\mathbf{x}^*} - \Sigma_{\mathbf{x}^*\mathbf{x}} \Sigma_{\mathbf{xx}}^{-1} \Sigma_{\mathbf{xx}^*} \,. \tag{A.5}$$

That is, given a weight u that follows a Gamma distribution with parameters ν^*:

$$u \sim Gamma(\nu^*/2, \nu^*/2) \,, \tag{A.6}$$

the vector $(\mathbf{x}^*, \mathbf{x})^T$ follows the multivariate normal distribution with mean μ^* and covariance Σ^*/u:

$$\mathbf{x}^*|\mathbf{x} \sim \mathcal{N}(\mu^*, \Sigma^*/u) \,. \tag{A.7}$$

From the properties of the t-distribution, it can be shown that, if $\nu^* > 1$, then μ^* is the mean of $(\mathbf{x}^*, \mathbf{x})^T$ and if $\nu^* > 2$, then $\nu^*(\nu^* - 2)^{-1}\Sigma^*$ is the covariance matrix of $\mathbf{x}^*$. Therefore, the family of t-distributions provides a heavy-tailed alternative to the normal family with mean μ and covariance matrix that is equal to a scalar multiple of Σ.

Appendix B

Learning Using Privileged Information

B.1 SVM+ Formulation

B.1 SVM+ Formulation

Learning using privileged information (LUPI) was originally introduced by Vapnik and Vashist [313]. Their SVM+ method is based on a max-margin classification scheme (SVM) and encodes additional (privileged) information about the training data, which is accessible only during training but never during testing. Many variants of SVM+ have been proposed, including SVM+ with $L1$ regularization [340], and multi-task SVM+ [341].

Given a training dataset with N samples $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{x}_i^*, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^{M_\mathbf{x}}$ is an observation sequence, which belongs in feature space $\mathcal{X}$, the privileged information is represented by $\mathbf{x}_i^* \in \mathbb{R}^{M_{\mathbf{x}^*}}$, which belongs to the privileged space $\mathcal{X}^*$, and y_i is the true class label defined in a finite label set $\mathcal{Y}$. The SVM+ algorithm determines the decision hyperplane between the two classes by parameterizing the slack variables ξ_i as a function of privileged features, $\xi_i = \langle \mathbf{w}^*, \mathbf{x}_i^* \rangle + b^*$, where $\mathbf{w}^*$ and b^* are the privileged parameters that are learned as a solution to the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{w},b,\mathbf{w}^*,b^*} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\|\mathbf{w}^*\|^2 + C\sum_{i=1}^{N}\xi_i\,, \\ &\text{subject to } \; y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \;\text{ and }\; \xi_i \geq 0, \;\; \forall i = 1, \ldots, N\,, \end{aligned} \tag{B.1}$$

where $\mathbf{w}$ is a normal vector perpendicular to the hyperplane, $C > 0$ and $\gamma > 0$ are two hyper-parameters that control the influence of the margin errors in the objective function, and b determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

For solving the optimization problem in Eq. (B.1) the standard technique is to consider its dual problem and construct the Lagrangian. Thus, the optimization problem becomes:

$$\begin{aligned} &\max_{\alpha,\beta} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{2\gamma} \sum_{i,j=1}^{N} (\alpha_i + \beta_i - C)(\alpha_j + \beta_j - C) K(\mathbf{x}_i^*, \mathbf{x}_j^*) \,, \\ &\text{subject to } \sum_{i=1}^{N} (\alpha_i + \beta_i - C) = 0, \;\; \sum_{i=1}^{N} y_i \alpha_i = 0, \;\; \alpha_i \geq 0, \;\; \beta_i \geq 0, \;\; \forall i = 1, \ldots, N \,, \end{aligned} \tag{B.2}$$

where α and β are the Lagrange dual variables of the SVM+, and $K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\mathbf{x}_i^*, \mathbf{x}_j^*)$ are kernel functions in the decision $\mathcal{X}$ and the correcting $\mathcal{X}^*$ space, respectively.

The decision function $f(\mathbf{x})$ takes place in the original space $\mathcal{X}$:

$$f(\mathbf{x}) = \operatorname{sgn} \sum_{i,j=1}^{N} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_j) \,. \tag{B.3}$$

Although only $K(\mathbf{x}_i, \mathbf{x}_j)$ contributes to the decision function, both $K(\mathbf{x}_i, \mathbf{x}_j)$ and $K(\mathbf{x}_i^*, \mathbf{x}_j^*)$ kernels are coupled through variable α in Eq. (B.2). In can be seen that Eq. (B.2) includes the solution to the standard SVM, therefore, SVM+ may either use privileged information only when it is considered to be informative by controlling the maximum influence of original space on the decision boundary or use the SVM solution instead.

Bibliography

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE International Conference on Computer Vision*, Beiging, China, 2005, pp. 1395–1402.

[2] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. International Conference on Pattern Recognition*, Cambridge, UK, 2004, pp. 32–36.

[3] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[4] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009, pp. 1–8.

[5] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Classifying behavioral attributes using conditional random fields," in *Proc. Hellenic Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science Volume 8445, Ioannina, Greece, May 2014, pp. 95–104.

[6] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in TV shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, December 2012.

[7] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, June 2012, pp. 28–35.

[8] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *Proc. European Conference on Computer Vision*, ser. Lecture Notes in Computer Science Volume 7575. Florence, Italy: Springer, October 2012, pp. 530–543.

[9] B. Schuller, M. Valstar, F. Eybenn, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011-the first international audio/visual emotion challenge," in *Proc. International Conference on Affective Computing and Intelligent Interaction - Volume Part II*. Memphis, TN, USA: Springer-Verlag, October 2011, pp. 415–424.

[10] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Computer Society Conference onComputer Vision and Pattern Recognition Workshop for Human Communicative Behavior Analysis*, San Francisco, CA, USA, June 2010, pp. 94–101.

[11] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[12] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density for visual surveillance," *Proc. of the IEEE*, vol. 90, no. 7, pp. 1151–1163, July 2002.

[13] A. Mumtaz, W. Zhang, and A. B. Chan, "Joint motion segmentation and background estimation in dynamic scenes," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 368–375.

[14] J. Liu, J. Yan, M. Tong, and Y. Liu, "A Bayesian framework for 3D human motion tracking from monocular image," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA, March 2010, pp. 1398–1401.

[15] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[16] X. Yan, I. A. Kakadiaris, and S. K. Shah, "Modeling local behavior for predicting social interactions towards human tracking," *Pattern Recognition*, vol. 47, no. 4, pp. 1626–1641, 2014.

[17] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 2847–2854.

[18] C. Gan, N. Wang, Y. Yang, D. Y. Yeung, and A. G. Hauptmann, "DevNet: A deep event network for multimedia event detection and evidence recounting," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 2568–2577.

[19] M. Jainy, J. C. Gemerty, and C. G. M. Snoek, "What do 15,000 object categories tell us about classifying and localizing actions?" in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 46–55.

[20] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1635–1648, 2013.

[21] B. Ni, P. Moulin, X. Yang, and S. Yan, "Motion part regularization: Improving action recognition via trajectory group selection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 3698–3706.

[22] K. N. Tran, A. Gala, I. A. Kakadiaris, and S. K. Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," *Pattern Recognition Letters*, vol. 44, pp. 49–57, 2014.

[23] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, July 2014.

[24] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1354–1361.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer, 2006.

[26] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning*, Williams College, Williamstown, MA, USA, 2001, pp. 282–289.

[27] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007.

[28] D. M. Gavrila, "The visual analysis of human movement: a survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[29] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, March 1999.

[30] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, 2003.

[31] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, November 2006.

[32] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *Proc. IEEE Transactions Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[33] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[34] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[35] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, November 2013.

[36] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yangg, and J. Gall, "A survey on human motion analysis from depth data," in *Proc. Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, ser. Lecture Notes in Computer Science Volume 8200, M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, Eds. Springer, 2013, pp. 149–187.

[37] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.

[38] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognition*, vol. 47, no. 10, pp. 3343–3361, 2014.

[39] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, pp. 116–134, 2007, special Issue on Vision for Human-Computer Interaction.

[40] M. Pantic and L. J. M. Rothkrantz, "Towards an affect-sensitive multimodal human-computer interaction," *Proc. IEEE, Special Issue on Multimodal Human-Computer Interaction, Invited Paper*, vol. 91, no. 9, pp. 1370–1390, September 2003.

[41] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: A survey," in *Proc. International Conference on Multimodal Interfaces*, New York, USA, November 2006, pp. 239–248.

[42] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[43] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image and Vision Computing*, vol. 31, no. 2, pp. 203–221, 2013.

[44] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A survey on ontologies for human behavior recognition," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–33, March 2014.

[45] A. H. Shabani, D. Clausi, and J. S. Zelek, "Improved spatio-temporal salient feature detection for action recognition," in *Proc. British Machine Vision Conference*, Dundee, UK, 2011, pp. 1–12.

[46] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 2855–2862.

[47] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1362–1369.

[48] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Action recognition by matching clustered trajectories of motion vectors," in *Proc. International Conference on Computer Vision Theory and Applications*, Barcelona, Spain, February 2013, pp. 112–117.

[49] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2003–2010.

[50] A. Iosifidis, A. Tefas, and I. Pitas, "Activity-based person identification using fuzzy representation and discriminant learning," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 530–542, 2012.

[51] V. I. Morariu and L. S. Davis, "Multi-agent event recognition in structured scenarios," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3289–3296.

[52] C. Y. Chen and K. Grauman, "Efficient activity detection with max-subgraph search," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1274–1281.

[53] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition," *Pattern Recognition*, vol. 45, no. 7, pp. 2562–2572, 2012.

[54] L. Sigal, M. Isard, H. Haussecker, and M. J. Black, "Loose-limbed people: estimating 3D human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 15–48, May 2012.

[55] Q. Wu, Z. Wang, F. Deng, Z. Chi, and D. D. Feng, "Realistic human action recognition with multimodal feature selection and fusion," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 4, pp. 875–885, 2013.

[56] N. Liu, E. Dellandréa, B. Tellez, and L. Chen, "Associating textual features with visual ones to improve affective image classification," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science Volume 6974, Memphis, TN, USA, October 2011, pp. 195–204.

[57] Y. Song, L. P. Morency, and R. Davis, "Multimodal human behavior analysis: learning correlation and interaction across modalities," in *Proc. ACM International Conference on Multimodal Interaction*, Santa Monica, CA, USA, 2012, pp. 27–30.

[58] M. J. Marín-Jiménez, R. M. noz Salinas, E. Yeguas-Bolivar, and N. P. de la Blanca, "Human interaction categorization by using audio-visual cues," *Machine Vision and Applications*, vol. 25, no. 1, pp. 71–84, 2014.

[59] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science Volume 4738, Lisbon, Portugal, 2007, pp. 71–82.

[60] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1775–1788, September 2014.

[61] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, and A. G. Hauptmann, "Semi-supervised multiple feature analysis for action recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 289–298, 2014.

[62] P. Matikainen, M. Hebert, and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features," in *Proc. Workshop on Video-Oriented Object and Event Classification, in conjunction with ICCV*, Kyoto, Japan, September 2009, pp. 514–521.

[63] M. Raptis, I. Kokkinos, and S. Soatto, "Discovering discriminative action parts from mid-level video representations," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1242–1249.

[64] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Matching mixtures of curves for human action recognition," *Computer Vision and Image Understanding*, vol. 119, no. 0, pp. 27–40, 2014.

[65] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, Nice, France, 2003, pp. 726–733.

[66] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[67] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[68] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[69] Y. Wang and G. Mori, "Hidden part models for human action recognition: probabilistic versus max margin," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1310–1323, 2011.

[70] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2287–2301, 2011.

[71] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. European Conference on Computer Vision*. Graz, Austria: Springer, 2006, pp. 428–441.

[72] R. Chaudhry, A. Ravichandran, G. D. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009, pp. 1932–1939.

[73] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, 2009, pp. 444–451.

[74] A. Oikonomopoulos, M. Pantic, and I. Patras, "Sparse B-spline polynomial descriptors for human activity recognition," *Image and Vision Computing*, vol. 27, no. 12, pp. 1814–1825, November 2009.

[75] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Sep. 2001.

[76] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[77] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. International Conference on Computer Communications and Networks*, Beijing, China, 2005, pp. 65–72.

[78] X. Yan and Y. Luo, "Recognizing human actions using a new descriptor based on spatial-temporal interest points and weighted-output classifier," *Neurocomputing*, vol. 87, pp. 51–61, 2012.

[79] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 2008, pp. 1–8.

[80] A. Yao, J. Gall, and L. V. Gool, "A Hough transform-based voting framework for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 2061–2068.

[81] K. Schindler and L. V. Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[82] I. Fogel and D. Sagi, "Gabor filters as texture discriminator," *Biological Cybernetics*, vol. 61, no. 2, pp. 103–113, Jun. 1989.

[83] M. Sapienza, F. Cuzzolin, and P. H. S. Torr, "Learning discriminative space-time action parts from weakly labelled videos," *International Journal of Computer Vision*, vol. 110, no. 1, pp. 30–47, 2014.

[84] S. Khamis, V. I. Morariu, and L. S. Davis, "A flow model for joint action recognition and identity maintenance," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1218–1225.

[85] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2005, pp. 886–893.

[86] J. Wang, Z. Chen, and Y. Wu, "Action recognition with multiscale spatio-temporal contexts," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3185–3192.

[87] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 2458–2466.

[88] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 2642–2649.

[89] M. Jain, J. Gemert, H. Jégou, P. Bouthemy, and C. G. M. Snoek, "Action localization with tubelets from motion," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 740–747.

[90] K. Kulkarni, G. Evangelidis, J. Cech, and R. Horaud, "Continuous action recognition based on sequence alignment," *International Journal of Computer Vision*, vol. 112, no. 1, pp. 90–114, 2015.

[91] S. Samanta and B. Chanda, "Space-time facet model for human activity classification," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1525–1535, October 2014.

[92] R. M. Haralick and L. Watson, "A facet model for image data," *Computer Graphics and Image Processing*, vol. 15, no. 2, pp. 113–129, 1981.

[93] Z. Jiang, Z. Lin, and L. S. Davis, "A unified tree-based framework for joint action localization, recognition and segmentation," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1345–1355, 2013.

[94] M. J. Roshtkhari and M. D. Levine, "Human activity recognition in videos using a single example," *Image and Vision Computing*, vol. 31, no. 11, pp. 864–876, 2013.

[95] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3361–3368.

[96] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1234–1241.

[97] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, September 2005.

[98] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proc. IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 2011, pp. 489–496.

[99] G. Yu, J. Yuan, and Z. Liu, "Propagative Hough voting for human activity recognition," in *Proc. European Conference on Computer Vision*. Florence, Italy: Springer, 2012, pp. 693–706.

[100] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 2555–2562.

[101] A. Gaidon, Z. Harchaoui, and C. Schmid, "Activity representation with motion hierarchies," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 219–238, 2014.

[102] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1302–1311.

[103] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier, "Activity recognition using dynamic subspace angles," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3193–3200.

[104] R. Messing, C. J. Pal, and H. A. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. IEEE International Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 104–111.

[105] D. Tran, J. Yuan, and D. Forsyth, "Video event detection: From subvolume localization to spatiotemporal path search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 404–416, 2014.

[106] M. B. Holte, B. Chakraborty, J. Gonzàlez, and T. B. Moeslund, "A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 553–565, 2012.

[107] Q. Zhou and G. Wang, "Atomic action features: A new feature for action recognition," in *Proc. European Conference on Computer Vision*. Firenze, Italy: Springer, 2012, pp. 291–300.

[108] J. Sanchez-Riera, J. Cech, and R. Horaud, "Action recognition robust to background clutter by using stereo vision," in *Proc. European Conference on Computer Vision*. Firenze, Italy: Springer, 2012, pp. 332–341.

[109] M. Hoai, Z. Z. Lan, and F. Torre, "Joint segmentation and classification of human actions in video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3265–3272.

[110] S. Satkin and M. Hebert, "Modeling the temporal extent of actions," in *Proc. European Conference on Computer Vision*. Heraklion, Crete, Greece: Springer, 2010, pp. 536–548.

[111] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann, "Action recognition by exploring data distribution and feature correlation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1370–1377.

[112] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396–410, 2012.

[113] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, August 2012.

[114] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 867–882, 2011.

[115] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 2046–2053.

[116] S. Ma, L. Sigal, and S. Sclaroff, "Space-time tree ensemble for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 5024–5032.

[117] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 5378–5387.

[118] N. Robertson and I. Reid, "A general method for human activity recognition in video," *Computer Vision and Image Understanding*, vol. 104, no. 2–3, pp. 232–248, 2006.

[119] Y. Wang and G. Mori, "Learning a discriminative hidden part model for human action recognition," in *Proc. Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2008, pp. 1721–1728.

[120] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, August 2000.

[121] Y. Song, L. P. Morency, and R. Davis, "Action recognition by hierarchical sequence summarization," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 3562–3569.

[122] Z. F. Huang, W. Yang, Y. Wang, and G. Mori, "Latent boosting for action recognition," in *Proc. British Machine Vision Conference*, Dundee, UK, 2011, pp. 1–11.

[123] S. Yi, H. Krim, and L. K. Norris, "Human activity as a manifold-valued random process," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3416–3428, 2012.

[124] C. Sun and R. Nevatia, "ACTIVE: activity concept transitions in video event classification," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 913–920.

[125] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[126] B. Ni, V. R. Paramathayalan, and P. Moulin, "Multiple granularity analysis for fine-grained action detection," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 756–763.

[127] T. Lan, T. C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *Proc. European Conference on Computer Vision*. Zurich, Switzerland: Springer, September 2014, pp. 689–704.

[128] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *Proc. European Conference on Computer Vision*. Zurich, Switzerland: Springer, September 2014, pp. 596–611.

[129] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3273–3280.

[130] S. J. D. Prince, *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.

[131] J. Lu, R. Xu, and J. J. Corso, "Human action segmentation with hierarchical supervoxel consistency," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 3762–3771.

[132] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 3169–3176.

[133] Z. Wang, J. Wang, J. Xiao, K. H. Lin, and T. S. Huang, "Substructure and boundary modeling for continuous action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1330–1337.

[134] M. R. Amer and S. Todorovic, "Sum-product networks for modeling activities with stochastic structure," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1314–1321.

[135] W. Zhou and Z. Zhang, "Human action recognition with multiple-instance Markov model," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1581–1591, 2014.

[136] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 748–755.

[137] Y. Kong and Y. Fu, "Modeling supporting regions for close human interaction recognition," in *Proc. European Conference on Computer Vision*. Zurich, Switzerland: Springer, September 2014, pp. 29–44.

[138] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 4576–4584.

[139] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-n-patch: Unsupervised understanding of actions and relations," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 4362–4370.

[140] C. Xu, S. H. Hsieh, C. Xiong, and J. J. Corso, "Can humans fly? Action understanding with multiple classes of actors," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 2264–2273.

[141] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 17–24.

[142] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, October 2009.

[143] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *Proc. European Conference on Computer Vision*, ser. Lecture Notes in Computer Science Volume 6311. Heraclion, Crete, Greece: Springer, 2010, pp. 494–507.

[144] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 3337–3344.

[145] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 780–787.

[146] H. Pirsiavash and D. Ramanan, "Parsing videos of actions with segmental grammars," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 612–619.

[147] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 3156–3164.

[148] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proc. International Joint Conference on Artificial Intelligence*, Barcelona, Spain, July 2011, pp. 1237–1242.

[149] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 2625–2634.

[150] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *Proc. European Conference on Computer Vision*. Heraklion, Crete, Greece: Springer, 2010, pp. 155–168.

[151] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 819–826.

[152] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2009, pp. 1410–1418.

[153] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, June 2009, pp. 951–958.

[154] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, November 2011, pp. 1331–1338.

[155] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Attribute regularization based human action recognition," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1600–1609, 2013.

[156] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, 2004, pp. 109–117.

[157] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Robust relative attributes for human action recognition," *Pattern Analysis and Applications*, vol. 18, no. 1, pp. 157–171, 2015.

[158] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei, "Learning semantic relationships for better action retrieval in images," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1100–1109.

[159] P. Kulkarni, G. Sharma, J. Zepeda, and L. Chevallier, "Transfer learning via attributes for improved on-the-fly classification," in *Porc. IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, USA, March 2014, pp. 220–226.

[160] R. Salakhutdinov, A. Torralba, and J. B. Tenenbaum, "Learning to share visual appearance for multiclass object detection." in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, June 2011, pp. 1481–1488.

[161] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, June 2009, pp. 1778–1785.

[162] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Proc. Annual Conference on Neural Information Processing Systems*, Montreal, Quebec, Canada, December 2014, pp. 3464–3472.

[163] C. Thurau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[164] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010, pp. 2030–2037.

[165] S. Maji, L. D. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3177–3184.

[166] S. Sedai, M. Bennamoun, and D. Q. Huynh, "Discriminative fusion of shape and appearance features for human pose estimation," *Pattern Recognition*, vol. 46, no. 12, pp. 3223–3237, 2013.

[167] I. Lillo, A. Soto, and J. C. Niebles, "Discriminative hierarchical modeling of spatio-temporally composable human activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 812–819.

[168] N. İkizler and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," in *Proc. Conference on Human Motion: understanding, modeling, capture and animation*, Rio de Janeiro, Brazil, 2007, pp. 271–284.

[169] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*, 4th ed. Boston, MA, USA: Academic Press, 2008.

[170] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, September 2012.

[171] T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds., *Self-Organizing Maps*, 3rd ed. Springer-Verlag New York, Inc., 2001.

[172] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.

[173] M. Andriluka and L. Sigal, "Human context: modeling human-human interactions for monocular 3D pose estimation," in *Proc. International Conference on Articulated Motion and Deformable Objects*. Mallorca, Spain: Springer-Verlag, 2012, pp. 260–272.

[174] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 3686–3693.

[175] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1290–1297.

[176] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 538–552, 2012.

[177] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 1297–1304.

[178] M. Fergie and A. Galata, "Mixtures of Gaussian process models for human pose estimation," *Image and Vision Computing*, vol. 31, no. 12, pp. 949–957, 2013.

[179] S. Sedai, M. Bennamoun, and D. Q. Huynh, "A Gaussian process guided particle filter for tracking 3D human pose in video," *IEEE Transactions on Image Processing*, vol. 22, no. 11, pp. 4286–4300, 2013.

[180] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view pictorial structures for 3D human pose estimation," in *Proc. British Machine Vision Conference*, Bristol, UK, September 2013, pp. 1–12.

[181] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic, "3D pictorial structures for multiple human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 1669–1676.

[182] L. Pishchulin, M. Andriluka, P. V. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 3487–3494.

[183] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 2337–2344.

[184] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 1653–1660.

[185] D. C. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012, pp. 3642–3649.

[186] H. Y. Jung, S. Lee, Y. S. Heo, and I. D. Yun, "Random treewalk toward instantaneous 3D human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 2467–2474.

[187] M. Livne, L. Sigal, N. F. Troje, and D. J. Fleet, "Human attributes from 3D pose tracking," *Computer Vision and Image Understanding*, vol. 116, no. 5, pp. 648–660, 2012.

[188] A. Moutzouris, J. M. del Rincon, J. C. Nebel, and D. Makris, "Efficient tracking of human poses using a manifold hierarchy," *Computer Vision and Image Understanding*, vol. 132, pp. 75–86, 2015.

[189] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1110–1118.

[190] B. X. Nie, C. Xiong, and S. C. Zhu, "Joint action recognition and pose estimation from video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1293–1301.

[191] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Pose search: Retrieving people using their pose," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, June 2009, pp. 1–8.

[192] V. K. Singh and R. Nevatia, "Action recognition in cluttered dynamic scenes using pose-specific part models," in *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, November 2011, pp. 113–120.

[193] A. Cherian, J. Mairal, K. Alahari, and C. Schmid, "Mixing body-part sequences for human pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 2361–2368.

[194] R. Xu, P. Agarwal, S. Kumar, V. N. Krovi, and J. J. Corso, "Combining skeletal pose with local motion for human activity recognition," in *Proc. International Conference on Articulated Motion and Deformable Objects*, Port d'Andratx, Mallorca, Spain, July 2012, pp. 114–123.

[195] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *Proc. Asian Conference on Computer Vision*, Singapore, November 2014, pp. 428–443.

[196] I. Kviatkovsky, E. Rivlin, and I. Shimshoni, "Online action recognition using covariance of shape and motion," *Computer Vision and Image Understanding*, vol. 129, pp. 15–26, 2014.

[197] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. S. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, USA, March 2014, pp. 626–633.

[198] T. K. Ho, "Random decision forests," in *Proc. IEEE International Conference on Document Analysis and Recognition*, vol. 1, Washington, DC, USA, August 1995, pp. 278–282.

[199] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 588–595.

[200] R. M. Murray, S. S. Sastry, and L. Zexiang, *A Mathematical Introduction to Robotic Manipulation*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 1994.

[201] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 3147–3155.

[202] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev, and J. Sivic, "People watching: Human actions as a cue for single view geometry," *International Journal of Computer Vision*, vol. 110, no. 3, pp. 259–274, 2014.

[203] R. Urtasun and T. Darrell, "Sparse probabilistic regression for activity-independent human pose inference," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 2008, pp. 1–8.

[204] I. Theodorakopoulos, D. Kastaniotis, G. Economou, and S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 12–23, 2014.

[205] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black, "Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation," *International Journal of Computer Vision*, vol. 98, no. 1, pp. 15–48, 2012.

[206] M. Burenius, J. Sullivan, and S. Carlsson, "3D pictorial structures for multiple view articulated pose estimation," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 3618–3625.

[207] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, 2015.

[208] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Los Alamitos, CA, USA, 2009, pp. 58–65.

[209] J. Lichtenauer, J. S. M. Valstar, and M. Pantic, "Cost-effective solution to synchronised audio-visual data capture using multiple sensors," *Image and Vision Computing*, vol. 29, no. 10, pp. 666–680, September 2011.

[210] P. Perez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proc. IEEE*, vol. 92, no. 3, pp. 495–513, March 2004.

[211] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[212] J. Healey, "Recording affect in the field: Towards methods and metrics for improving ground truth labels," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, Memphis, TN, USA, October 2011, pp. 107–116.

[213] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[214] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[215] A. Klami and S. Kaski, "Probabilistic approach to detecting dependencies between data sets," *Neurocomputing*, vol. 72, no. 1-3, pp. 39–46, 2008.

[216] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton University Press, 1976.

[217] M. S. Hussain, R. A. Calvo, and P. A. Pour, "Hybrid fusion approach for detecting affects from multichannel physiology," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science Volume 6974, Memphis, TN, USA, October 2011, pp. 568–577.

[218] O. AlZoubi, D. Fossati, S. K. D'Mello, and R. A. Calvo, "Affect detection and classification from the non-stationary physiological data," in *Proc. International Conference on Machine Learning and Applications*, Portland, OR, USA, December 2013, pp. 240–245.

[219] B. Siddiquie, S. M. Khan, A. Divakaran, and H. S. Sawhney, "Affect analysis in natural human interaction using joint hidden conditional random fields," in *Proc. IEEE International Conference on Multimedia and Expo*, San Jose, CA, USA, July 2013, pp. 1–6.

[220] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011 -the first international audio visual emotion challenge," in *Proc. International Audio/Visual Emotion Challenge and Workshop*, ser. Lecture Notes in Computer Science Volume 6975, Memphis, TN, USA, 2011, pp. 415–424.

[221] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April 2011.

[222] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, August 2004.

[223] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. International Conference on Machine Learning*, Bellevue, WA, USA, 2011, pp. 689–696.

[224] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, May 2013.

[225] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper, and R. Kasturi, "Understanding transit scenes: a survey on human behavior-recognition algorithms," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 1, pp. 206–224, 2010.

[226] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Shanghai, China, April 2013, pp. 1–8.

[227] M. E. Sargin, Y. Yemez, E. Erzin, and A. M.Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[228] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor., "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, December 2004.

[229] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in *Proc. IEEE International Symposium on Multimedia*, Berkeley, CA, USA, December 2008, pp. 250–257.

[230] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System (FACS): Manual*. Salt Lake City, UT, USA: A Human Face, 2002.

[231] H. Chen, J. Li, F. Zhang, Y. Li, and H. Wang, "3D model-based continuous emotion recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1836–1845.

[232] Q. Wu, Z. Wang, F. Deng, and D. D. Feng, "Realistic human action recognition with audio context," in *Proc. International Conference on Digital Image Computing: Techniques and Applications*, Sydney, Australia, December 2010, pp. 288–293.

[233] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 2008, pp. 1–8.

[234] Y. Song, L. P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012, pp. 2120–2127.

[235] A. Metallinou, M. Wollmer, A. Katsamani, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, April 2012.

[236] K. Bousmalis, S. Zafeiriou, L. P. Morency, and M. Pantic, "Infinite hidden conditional random fields for human behavior analysis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 170–177, 2013.

[237] R. H. Baxter, N. M. Robertson, and D. M. Lane, "Human behaviour recognition in data-scarce domains," *Pattern Recognition*, vol. 48, no. 8, pp. 2377–2393, 2015.

[238] G. Rawlinson, "The significance of letter position in word recognition," *IEEE Aerospace and Electronic Systems Magazine*, vol. 22, no. 1, pp. 26–27, January 2007.

[239] S. Bilakhia, S. Petridis, and M. Pantic, "Audiovisual detection of behavioural mimicry," in *Proc. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, September 2013, pp. 123–128.

[240] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[241] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. British Machine Vision Conference*, University of Leeds, Leeds, UK, September 2008, pp. 995–1004.

[242] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1226–1233.

[243] H. S. Park and J. Shi, "Social saliency prediction," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 4777–4785.

[244] X. P. Burgos-Artizzu, P. Dollár, D. Lin, D. J. Anderson, and P. Perona, "Social behavior recognition in continuous video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 1322–1329.

[245] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3161–3167.

[246] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Learning multimodal latent attributes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 303–316, 2014.

[247] W. L. Lu, J. A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3249–3256.

[248] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–1562, August 2012.

[249] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Miami Beach, FL, USA, June 2009, pp. 2929–2936.

[250] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic, "Finding actors and actions in movies," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 2280–2287.

[251] M. Hoai and A. Zisserman, "Talking heads: Detecting humans and recognizing their interactions," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 875–882.

[252] V. Ramanathan, P. Liang, and L. Fei-Fei, "Video event understanding using natural language descriptions," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 905–912.

[253] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. J. Mooney, T. Darrell, and K. Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 2712–2719.

[254] S. Bandla and K. Grauman, "Active learning of an action detector from untrimmed videos," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 1833–1840.

[255] K. Bousmalis, L. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, CA, USA, March 2011, pp. 746–752.

[256] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1766–1778, 2014.

[257] S. Escalera, X. Baró, J. Vitrià, P. Radeva, and B. Raducanu, "Social network extraction and analysis based on multimodal dyadic interaction," *Sensors*, vol. 12, no. 2, pp. 1702–1719, 2012.

[258] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013, pp. 3687–3691.

[259] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[260] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 4657–4666.

[261] Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in *Proc. IEEE Computer Society Conference on*

Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015, pp. 1600–1609.

[262] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 1725–1732.

[263] L. Chen, L. Duan, and D. Xu, "Event recognition in videos by learning from heterogeneous web sources," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 2666–2673.

[264] K. D. Tang, B. Yao, L. Fei-Fei, and D. Koller, "Combining the right features for complex event recognition," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 2696–2703.

[265] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 2211–2218.

[266] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[267] S. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey," *Proc. IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[268] Q. S. Sun, S. G. Zeng, Y. Liu, P. A. Heng, and D. S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, Dec. 2005.

[269] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for multimodal information fusion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2384–2387.

[270] O. Rudovic, S. Petridis, and M. Pantic, "Bimodal log-linear regression for fusion of audio and visual features," in *Proc. ACM Multimedia Conference*, Barcelona, Spain, October 2013, pp. 789–792.

[271] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, November 2013.

[272] T. Westerveld, A. P. de Vries, A. van Ballegooij, F. de Jong, and D. Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processecing*, vol. 2003, pp. 186–198, Jan. 2003.

[273] B. Jiang, B. Martínez, M. F. Valstar, and M. Pantic, "Decision level fusion of domain specific regions for facial action recognition," in *Proc. International Conference on Pattern Recognition*, Stockholm, Sweden, August 2014, pp. 1776–1781.

[274] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proc. Annual ACM International Conference on Multimedia*, Singapore, 2005, pp. 399–402.

[275] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[276] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*, Boston, MA, USA, 2010, pp. 48–55.

[277] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[278] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR, abs/1212.0402*, November 2012.

[279] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, 2011, pp. 2556–2563.

[280] R. B. Fisher, "PETS04 surveillance ground truth dataset," http://www-prima.inrialpes.fr/PETS04/, Prague, Czech Republic, May 2004.

[281] R. B. Fisher, "PETS07 benchmark dataset," http://www.cvg.reading.ac.uk/PETS2007/data.html, Rio de Janeiro, Brazil, October 2007.

[282] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 3153–3160.

[283] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Proc. IEEE*

International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), Kyoto, Japan, 2009, pp. 1089–1096.

[284] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proc. Conference on Human Factors in Computing Systems*, Austin, TX, USA, May 2012, pp. 1737–1746.

[285] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 3192–3199.

[286] M. Rohrbach, S. Amin, A. Mykhaylo, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 2012, pp. 1194–1201.

[287] Y. G. Jiang, G. Ye, S. F. Chang, D. P. W. Ellis, and A. C. Loui, "Consumer video understanding: a benchmark database and an evaluation of human and machine performance," in *Proc. International Conference on Multimedia Retrieval*, Trento, Italy, April 2011, pp. 29–36.

[288] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 961–970.

[289] R. B. Fisher, "Behave: Computer-assisted prescreening of video streams for unusual activities," http://homepages.inf.ed.ac.uk/rbf/BEHAVE/, 2007.

[290] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proc. International Conference on Affective Computing and Intelligent Interaction and Workshops*, De Rode Hoed Amsterdam, Netherlands, September 2009, pp. 1–4.

[291] A. Metallinou, C. C. Lee, C. Busso, S. M. Carnicke, and S. Narayanan, "The USC Creative IT database: A multimodal database of theatrical improvisation," in *Proc. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality.* Malta: Springer, May 2010, pp. 1–4.

[292] M. Vlachos, D. Gunopoulos, and G. Kollios, "Discovering similar multidimensional trajectories," in *Proc. International Conference on Data Engineering*, San Jose, CA, USA, 2002, pp. 673–682.

[293] F. Zhou and F. D. la Torre, "Canonical time warping for alignment of human behavior," in *Proc. Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2009, pp. 2286–2294.

[294] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[295] V. Karavasilis, K. Blekas, and C. Nikou, "Motion segmentation by model-based clustering of incomplete trajectories," in *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2011, pp. 146–161.

[296] V. Karavasilis, K. Blekas, and C. Nikou, "A novel framework for motion segmentation and tracking by clustering incomplete trajectories," *Computer Vision and Image Understanding*, vol. 116, no. 11, pp. 1135–1148, November 2012.

[297] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. International Joint Conference on Artificial Intelligence*, Nice, France, 1981, pp. 674–679.

[298] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, pp. 773–795, 1995.

[299] J. Kuha, "AIC and BIC: Comparisons of assumptions and performance," *Sociological Methods Research*, vol. 33, no. 2, pp. 188–229, November 2004.

[300] I. T. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[301] G. Sfikas, C. Constantinopoulos, A. Likas, and N. P. Galatsanos, "An analytic distance metric for gaussian mixture models with application in image retrieval," in *Proc. International Conference on Artificial Neural Networks*, Warsaw, Poland, September 2005, pp. 835–840.

[302] J. Domke, "Learning graphical model parameters with approximate marginal inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2454–2467, 2013.

[303] N. Komodakis and G. Tziritas, "Image completion using efficient belief propagation via priority scheduling and dynamic pruning," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2649–2661, November 2007.

[304] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[305] C. B. Germain and M. Bloom, *Human Behavior in the Social Environment: An Ecological View*. Columbia University Press, 1999.

[306] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed., ser. Springer series in operations research and financial engineering. New York, NY: Springer, 2006.

[307] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412–424, March 2006.

[308] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[309] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, February 2013.

[310] A. Gaidon, Z. Harchaoui, and C. Schmid, "Recognizing activities with cluster-trees of tracklets," in *Proc. British Machine Vision Conference*, Surrey, UK, September 2012, pp. 1–13.

[311] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang, "Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1553–1566, December 2004.

[312] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: an experimental survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1–1, 2014.

[313] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5–6, pp. 544–557, 2009.

[314] J. Chen, X. Liu, and S. Lyu, "Boosting with side information," in *Proc. Asian Conference on Computer Vision*, ser. Lecture Notes in Computer Science, vol. 7724, Daejeon, Korea, November 2012, pp. 563–577.

[315] J. Feyereisl and U. Aickelin, "Privileged information for data clustering," *Information Sciences*, vol. 194, no. 0, pp. 4–23, 2012.

[316] Z. Wang and Q. Ji, "Classifier learning with hidden information," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 4969–4977.

[317] V. Sharmanska, N. Quadrianto, and C. H. Lampert, "Learning to rank using privileged information," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 825–832.

[318] D. Peel and G. J. Mclachlan, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, pp. 339–348, 2000.

[319] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, December 2004.

[320] C. H. Teo, A. J. Smola, S. V. N. Vishwanathan, and Q. V. Le, "A scalable modular convex solver for regularized risk minimization," in *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, August 2007, pp. 727–736.

[321] S. Kotz and S. Nadarajah, *Multivariate t distributions and their applications*. Cambridge, New York, Madrid: Cambridge University Press, 2004.

[322] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[323] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 3551–3558.

[324] Z. Wang, T. Gao, and Q. Ji, "Learning with hidden information using a max-margin latent variable model," in *Proc. International Conference on Pattern Recognition*, Stockholm, Sweden, August 2014, pp. 1389–1394.

[325] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[326] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, March 2002.

[327] G. Druck, B. Settles, and A. McCallum, "Active learning by labeling features," in *Proc. Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 81–90.

[328] C. Constantinopoulos and A. Likas, "Semi-supervised and active learning with the probabilistic RBF classifier," *Neurocomputing*, vol. 71, no. 13–15, pp. 2489–2498, 2008.

[329] M. Hasan and A. K. Roy-Chowdhury, "Incremental activity modeling and recognition in streaming videos," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 796–803.

[330] S. J. Huang, S. Chen, and Z. H. Zhou, "Multi-label active learning: Query type matters," in *Proc. International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, July 2015, pp. 946–952.

[331] M. Hasan and A. K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, December 2015, pp. 4543–4551.

[332] C. Long, G. Hua, and A. Kapoor, "Active visual recognition with expertise estimation in crowdsourcing," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 3000–3007.

[333] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, October 2015.

[334] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor, "Exploiting sparsity and co-occurrence structure for action unit recognition," in *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[335] S. Parameswaran and K. Q. Weinberger, "Large margin multi-task metric learning," in *Proc. Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2010, pp. 1867–1875.

[336] R. Gopalan, L. Ruonan, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *Proc. IEEE International Conference on Computer Vision*, Barcelona, Spain, November 2011, pp. 999–1006.

[337] C. Serra-Toro, V. J. Traver, and F. Pla, "Exploring some practical issues of svm+: Is really privileged information that helps?" *Pattern Recognition Letters*, vol. 42, no. 0, pp. 40–46, 2014.

[338] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. IEEE International Conference on Multimedia and Expo*, Singapore, July 2010, pp. 1079–1084.

[339] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, June 2001.

[340] L. Niu, Y. Shi, and J. Wu, "Learning using privileged information with L-1 support vector machine," in *Proc. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 3, Macau, China, December 2012, pp. 10–14.

[341] Y. Ji and S. Sun, "Multitask multiclass support vector machines: Model and experiments," *Pattern Recognition*, vol. 46, no. 3, pp. 914–924, 2013.

Author's Publications

Publications related to this Thesis

Journal Publications

1. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **Human activity recognition using robust privileged probabilistic learning**, November, 2015 (submitted)

2. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **Identifying human behaviors using synchronized audio-visual cues**, IEEE Transactions on Affective Computing, 2016 (in press)

3. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **A review of human activity recognition methods**, Frontiers in Robotics and Artificial Intelligence, vol 2, no. 28, pp. 1-26, November 2015.

4. M. Vrigkas, V. Karavasilis, C. Nikou and I. A. Kakadiaris, **Matching mixtures of curves for human action recognition**, Computer Vision and Image Understanding, Vol. 119, pp. 27-40, February 2014.

Conference Publications

1. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **Exploiting privileged information for facial expression recognition**, January, 2016 (submitted)

2. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **Active privileged learning of human activities from weakly labeled samples**, January, 2016 (submitted)

3. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **Inferring human activities using robust privileged probabilistic learning**, November, 2015 (submitted)

4. M. Vrigkas, C. Nikou and I. A. Kakadiaris, **Classifying behavioral attributes using conditional random fields**, in Proc. 8th Hellenic Conference on Artificial Intelligence, ser. Lecture Notes in Computer Science, vol. 8445, pp. 95-104, Ioannina, Greece, May 15-17 2014.

5. M. Vrigkas, V. Karavasilis, C. Nikou and I. A. Kakadiaris, **Action recognition by matching clustered trajectories of motion vectors**, in Proc. 8th International Conference on Computer Vision Theory and Applications, pp. 112-117, Barcelona, Spain, February 21-24 2013.

Publications not related to this Thesis

Journal Publications

1. M. Vrigkas, C. Nikou and L. P. Kondi, **Robust maximum a posteriori image super-resolution**, Journal of Electronic Imaging, vol. 23, no. 4, pp. 043016, July 2014.

2. M. Vrigkas, C. Nikou and L. P. Kondi, **Accurate image registration for MAP image super-resolution**, Signal Processing: Image Communication, vol. 28, no. 5, pp. 494-508, May 2013.

Conference Publications

1. M. E. Plisiti, M. Vrigkas and C. Nikou, **Segmentation of cell clusters in Pap smear images using intensity variation between superpixels**, in Proc. 22nd International Conference on Systems, Signals and Image Processing, pp. 184-187, London, UK, September 10-12 2015.

2. M. Vrigkas, C. Nikou and L. P. Kondi, **A fully robust framework for MAP image super-resolution**, in Proc. IEEE International Conference on Image Processing, pp. 2225-2228, Orlando, FL, USA, September 30-October 3 2012.

3. M. Vrigkas, C. Nikou and L. P. Kondi, **On the improvement of image registration for high accuracy super-resolution**, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 981-984, Prague, Czech Republic, May 22-27 2011.

Short Vita

Michalis Vrigkas was born in Ioannina, Greece. He received the B.Sc. and M.Sc. in Computer Science from the University of Ioannina, Greece, in 2008 and 2010, respectively. Since 2011 he has been a Ph.D. student in the Department of Computer Science and Engineering, University of Ioannina, Greece. He is a member of the Information Processing and Analysis Research Group (I.PAN.). Since 2009, he has been working as a freelance and he has been involved in several national and EU-funded ICT projects. He is an IEEE member and since January 2015 he has been the Chair of the IEEE Student Branch of the University of Ioannina, Greece. His research interests include image processing and analysis, computer vision, machine learning and pattern recognition.