Robust Incremental Hidden Conditional Random Fields for Action Recognition

A Thesis

submitted to the designated
by the General Assembly of Special Composition
of the Department of Computer Science and Engineering
Examination Committee

by

Ermioni Mastora

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

WITH SPECIALIZATION
IN TECHNOLOGIES - APPLICATIONS

University of Ioannina
February 2017

Examining Committee:

- **Χριστόφορος Νίκου**, Αναπλ. Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων (Επιβλέπων)
- **Αριστείδης Λύκας**, Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
- **Κωνσταντίνος Μπλέκας**, Αναπλ. Καθηγητής, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων

Dedication

Paulo Coelho once wrote:
When you want something,
all the universe conspires in helping you to achieve it.

To the person who helps me realize my dreams and became my own universe ...

Acknowledgements

First of all, I would like to express my sincerest gratitude and thanks to my supervisor, Prof. Christoforos Nikou. His guidance, continued support, encouragement, invaluable insights and thoughtful suggestions helped me complete this research. I am honored to collaborate with him and I will always admire the superior dedication that he shows to his students.

Also, I would like to thank my fellow colleagues in the computer science department: Michalis Vrigkas and Evangelos Kazakos, who contributed significantly throughout the course of this project. I am extremely thankful to Michalis for his support, his immediate response to all my queries and for the fusion feature representation he provided to me. I thank Evangelos who has collected the C3D features for some tested datasets.

Additionally, I want to show my deepest gratitude to my dear friends and especially Dimitrios Zavantis, who have helped me in many ways and supported me to get through the hard time.

Last but not least, I would like to thank my family for their consistent support, advises, confidence and belief in me. Their constant emphasis on education has been an impact factor in the completion of this work and a motivation throughout my life.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Abstract

Ermioni Mastora, M.Sc. in Computer Science, Department of Computer Science and Engineering, University of Ioannina, Greece, February 2017.
Robust Incremental Hidden Conditional Random Fields for Action Recognition.
Advisor: Christoforos Nikou, Associate Professor.

Human action recognition is a challenging topic of computer vision research and continues to receive a keen interest due to the variety of applications that can be used. The creation of a supervised system able to understand and automatically recognize low-level actions and high-level activities is the core problem that these applications attempt to solve. A promising probabilistic graphical model that has been recently proposed for the recognition task is Hidden Conditional Random Fields (HCRF). However, the number of hidden variables that the model incorporates remains a severe limitation of the HCRF due to the fact that the user is asked to make an advance and intuitive assumption for this parameter.

In this thesis, we address this limitation by proposing a new model, called Robust Incremental Hidden Conditional Random Fields (RI-HCRF), which estimates the number of hidden states incrementally. Multiple Hidden Markov Models (HMM) are created whose parameters are defined by the potentials of the original HCRF graph. Starting from a small number of hidden states and increasing their number incrementally, the Viterbi path is computed for each HMM. The method seeks for a sequence of hidden states, where each variable participates in a maximum number of optimal paths. Therefore, variables with low participation in optimal paths are rejected. In addition, a robust mixture of Student's t-distributions is imposed as a regularizer to the parameters of the model.

The proposed method is tested in six publicly available datasets using different feature representations. The a priori knowledge of the optimal number of hidden variables and the t-distributed parameters lead to a more robust estimation framework for the classification task. The experiment results show that RI-HCRF estimates successfully the number of hidden states and outperforms all state-of-the-art models that were used as baseline.

Εκτεταμενη Περιληψη

Ερμιόνη Μάστορα, Μ.Δ.Ε. στην Πληροφορική, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Φεβρουάριος 2017.
Εύρωστα, Αυξητικά, Κρυφά Στοχαστικά Υπό Συνθήκη Πεδία για Αναγνώριση Δραστηριότητας.
Επιβλέπων: Χριστόφορος Νίκου, Αναπληρωτής Καθηγητής.

Το πρόβλημα της αναγνώρισης ανθρώπινης κίνησης παραμένει μια μεγάλη πρόκληση και αποτελεί ένα αρκετά ενεργό θέμα έρευνας για το πεδίο της μηχανικής όρασης. Η οπτική ανάλυση του περιεχομένου των εικονοσειρών κεντρίζει το ενδιαφέρον πολλών ερευνητών καθώς διαθέτει ένα μεγάλο εύρος εφαρμογών. Οι εφαρμογές αυτές περιλλαμβάνουν: συστήματα παρακολούθησης και καταγραφής εικόνας, ανάλυση αθλητικών βίντεο, συστήματα υγειονομικής περίθαλψης, αλληλεπίδραση ανθρώπου-ρομπότ είτε ανθρώπου-υπολογιστή και πολλές ακόμη. Ο στόχος τους είναι η δημιουργία ενός συστήματος το οποίο είναι σε θέση να κατανοεί και αναγνωρίζει αυτόματα ενέργειες χαμηλού επιπέδου καθώς και υψηλού επιπέδου δραστηριότητες. Ωστόσο, η αναγνώριση πολύπλοκων ανθρώπινων δραστηριοτήτων στον πραγματικό κόσμο είναι μία δύσκολη διαδικασία λόγω της ομοιότητας κάποιων κινήσεων, των μεταβολών στο φόντο, της φωτεινότητα, της κλίμακα είτε της μερικής εμφάνισης των εικονιζόμενων ατόμων.

Το πλήθος και η διάσταση των εικονοσειρών που είναι πλέον διαθέσιμα τα τελευταία χρόνια είναι αρκετά μεγάλα και η περιληπτική αναπαράσταση τους είναι πλέον απαραίτητη. Συνεπώς, οι εικονοσειρές θεωρούνται ως μία συλλογή από τοπικά χωροχρονικά χαρακτηριστικά. Διαχρονικά έχουν προταθεί πολλά μοντέλα για την αναγνώριση κίνησης αλλά πρόσφατα η έρευνα έχει στραφεί στην χρήση και τη δημιουργία νέων γραφικών μοντέλων. Τα κρυφά υπό συνθήκη τυχαία πεδία αποτελούν ένα πιθανοτικό μοντέλο όπου οι εξαρτήσεις μεταξύ των χωροχρονικών χαρακτηριστικών μπορούν να αποτυπωθούν και να αναπαρασταθούν υπό την μορφή ενός γράφου. Το μοντέλο αυτό έχει επιτύχει μεγάλη αύξηση στο ποσοστό επιτυχίας πολλών συνόλων δεδομένων σε σχέση με παλιότερα μοντέλα όμως, έχει ένα βασικό μειονέκτημα. Ο καθορισμός του αριθμού των κρυμμένων καταστάσεων, όπου το μοντέλο περιλαμβάνει, είναι μία παράμετρος και ζητείται από τον χρήστη να την καθορίσει συνήθως ενστικτωδώς εκ των προτέρων.

Στόχος αυτής της εργασίας είναι η εξάλειψη αυτού του μειονεκτήματος προτείνοντας ένα νέο μοντέλο, που ονομάζεται εύρωστα, αυξητικά, κρυφά, στοχαστικά υπό συνθήκη πεδία, η οποία προσθέτει στα κρυφά υπό συνθήκη πεδία μία αυξητική μέθοδο για την εκτίμηση του αριθμού των κρυμμένων καταστάσεων του μοντέλου. Για τον καθορισμό των κρυμμένων καταστάσεων δημιουργούνται πολλαπλά κρυμμένα Μαρκοβιανά μοντέλα όπου οι παράμετροι τους ορίζονται χρησιμοποιώντας τις συναρτήσεις του γραφήματος των κρυφών υπό συνθήκη πεδίων. Ξεκινώντας από ένα μικρό αριθμό κρυφών καταστάσεων και αυξάνοντας τον αριθμό τους σταδιακά, το βέλτιστο μονοπάτι (Viterbi) υπολογίζεται για κάθε Μαρκοβιανό μοντέλο. Η μέθοδος επιδιώκει μια αλληλουχία των κρυφών καταστάσεων, όπου κάθε κατάσταση συμμετέχει σε μεγάλο πλήθος βέλτιστων μονοπατιών. Ως εκ τούτου, οι καταστάσεις με χαμηλή συμμετοχή στα βέλτιστα μονοπάτια απορρίπτονται. Επιπλέον, η εύρωστη μεικτή κατανομή Student *t* προστίθεται στο μοντέλο ως την κατανομή που ακολουθούν οι παράμετροι του.

Η απόδοση της προτεινόμενης μεθόδου έχει εκτιμηθεί σε έξι σύνολα δεδομένων όπου για κάθε σύνολο χρησιμοποιήθηκαν διαφορετικά χωροχρονικά χαρακτηριστικά για την αναπαράστασή τους. Η προτεινόμενη μέθοδος εξετάζεται στην αναγνώριση χειρονομιών, κινήσεων, δραστηριοτήτων, εκδηλώσεων και κατηγοριών συμπεριφοράς. Ως αποτέλεσμα, συμπαιραίνουμε πως η εκ των προτέρων γνώση του βέλτιστου αριθμού των κρυμμένων καταστάσεων και οι Student *t* κατανεμημένες παράμετροι οδήγησαν σε ένα πιο εύρωστο μοντέλο ταξινόμησης. Τα αποτελέσματα των πειραμάτων δείχνουν ότι το μοντέλο μας προσδιορίζει με επιτυχία τον αριθμό των κρυμμένων καταστάσεων και ξεπερνά την απόδοση όλων των μοντέλων που χρησιμοποιήθηκαν για σύγκριση.

Chapter 1

Introduction

1.1 Research Background

The human visual system with the presence of light rays has the capability to assimilate, process and interpret information from its surrounding environment. Our visual perception is able to shape the physical world and recognize complex concepts, actions and interactions. Human action recognition has been an active and important area developed in the field of computer vision. The term action refers to an activity that can be a composition of simple movements (gestures) of body parts, which illustrate the motion of a person. Action recognition pertains to the study and development of models that can classify an action. The motivation of this area is to duplicate the ability of the human eye and map a label or automatically recognize the performed action in a video sequences.

A tremendous amount of human actions-video recordings has been generated, uploaded or stored in recent years. As a consequence, vision-based action recognition has become a challenging and applicable task to many types of ongoing actions. Visual surveillance systems, behavioral biometrics, human-computer interaction, video retrieval and sports video analysis are some important instances of these applications. Therefore, the implementation of action recognition algorithms is a significant impact factor in real world, demonstrated by the growing body of research these years.

The recognition of human activities is a challenging task due to the degree of intra and inter-class variations between the multimedia contents. Intra-class differences indicate anthropometric differences (size, gender, shape) among subjects and variations in the way, spread and speed of the action. Inter-class differences make reference to actions which do not differ much such as walking and jogging. Additionally, factors which advance the complication of this task may be considered such as record settings, background, type of audio-visual data, view-point, partial occlusion and lighting effects. As a result, the development of a general classification technique, which can be invariant to these variations, has received a significant attention in vision research.

There exist an extensive amount of action recognition techniques in the literature. The common framework they share is the use of features extracted from videos, for the recognition task. Features can reduce the complexity and dimension, sometimes are robust to noise, can discard the unnecessary information and can be rich descriptors over intra and inter class variations. According to Poppe's survey [7], action recognition methods, based on the feature representation they use, may be classified into two main categories: global and local.

Global methods based on holistic representation of an action [2], treat a video sequence or an image as one entity. Widely applied representations are silhouettes, optical flow or gradient, edges and motion trajectory. In holistic representations, spatio-temporal features are learned directly from sequential frames in a video. As a consequence of their ability to exploit large amounts of visual information and simultaneously preserve spatial and temporal structures of an ongoing action in a video, they have recently received significant attention [8–10]. However, holistic representations often require background segmentation and body tracking because of their sensitivity to occlusions and variations. Therefore, they need a pre-processing step, which is computationally expensive and impracticable in the majority of scenarios.

Local methods based on local representation of an action, use spatio-temporal interest points (STIPs) to express an action instance. The collection of local descriptors or local spatio-temporal local patches (features) extracted by STIPs represent the whole video sequence. This representation is less sensitive to occlusions and variations than the holistic one. In real-world scenes, it can be considered as state-of-the-art performance for action recognition tasks while it is combined with a bag-of-word (BOW) representation. Laptev *et al.* [11] proposed an extension of 2-D Harris corner detector to 3-D through the addition of time dimension. As interest points were indicated those where the image values showed significant variations in all dimensions. Interest points are usually described by histograms of gradients (HOG) and histograms of optical flows (HOF). There is a huge number of research for the estimation interest points. Among them, Dollar *et al.* [12] used temporally Gabor filters, Rapantzikos *et al.* [13] used discrete wavelet transform in temporal and spatial direction of the video and Scovanner *et al.* [14] extended 2D SIFT descriptor to 3D.

1.2 Motivation and Objectives

The design and selection of features for image and video representation have received a lot of intention and an extensive research in the literature. Nowadays, there is a large portfolio of features that serves a variety of applications and capable to fulfill different kind of application's demands. These features may be simple or complex, adopted by online or offline mode techniques, and used universally or for a specific application domain. Therefore, a huge collection of feature has been proposed over the years.

On the other hand, the number of proposed action recognition systems is not comparable to the feature's detectors. At the present time, methods like k-Nearest Neighbor (kNN) [15] and Support Vector Machines (SVM) [16] are still widely used because of their effectiveness and simplicity. However, these models are not designed for sequential data and can not encode into their structure dependencies among the features of the same action. Consequently, the problem that we address in this thesis is learning action categories from supervised data. We employ Hidden Conditional Random Fields (HCRF) [17], as a prominent and advanced classification method suitable to this task.

HCRF is a generalization of Conitional Random Fields (CRF) [18]. It is a powerful discriminative classification model, which engages into its structure hidden variables. These hidden variables have been proved to improve the performance of the model [19, 20] and are able to capture the motion pattern of an action. HCRF learn not only the hidden variable that discriminate one action category from all the others, but also models the spatial structures (dependencies) of image parts and temporal structures of video frames. Additionally, it allows the features of sequential frames to be overlapping and dependent on each other. As a result, it relaxes the restrictive and unrealistic independence assumption, which was used by other models.

However, like all models HCRF has its limitations. The number of hidden variables is not available and need to be fixed in advance. Setting an intuitive number of hidden variables with respect to the number of action categories or human poses during the performance of the action is not always correct. The common technique to this kind of problem, is to run the model trying different numbers of hidden variables and use a cross validation criterion to define the optimal one. Though, a technique like that is computationally expensive.

The objective of this thesis is to design and implement a technique that automatically estimates the optimal number of hidden variables in HCRF given a specific dataset. Along with this, in order to improve the recognition rate of the HCRF model we assume that the parameters of the model follow a mixture of Student's *t* distribution. The Student's *t* distribution is more robust to outliers compared to the Gaussian distribution that was originally proposed and can be exploited to improve the classification performance in the presence of abnormal values.

1.3 Contributions

The main contribution of this work is two-fold.

First, the main contribution comes in the form of a novel machine learning model for action recognition. We designed and implemented a novel, robust and incremental hidden conditional random field model. Driven by our goal to find a solution to the limitation of the estimation of the hidden variables, we propose an improved and extended version of the model. Based on the given observation sequences, our model is able to define in advance the optimal number of hidden variables using a Viterbi-like procedure. As a consequence, the apriori knowledge of the hidden variables is essential for the model to avoid a computationally expensive and trial-and-error process. Moreover, the model parameters follow a mixture of Student's *t* distribution, which makes the model robust to outliers. This adjustment has improved the performance of our model. We call the proposed model robust incremental hidden conditional random field (RI-HCRF).

Second, we applied our RI-HCRF action recognition system to six publicly available and challenging dataset. We showed that the design of our system is well suited to human - human interactions and the recognition performance of our system yielded a remarkable improvement. The layer of hidden variables, that interlace through time, is able to learn the relationships of the interaction and enable the model to comprehend the nature of these activities. One of the datasets reached for the first time a recognition accuracy rate of 100% while the rest of datasets achieved a recognition rate greater or equal to 89%. The model was applied using different feature configurations.

1.4 Thesis outline

This thesis is organized as follows:

In Chapter 2, the background and the related work for recognition systems in literature is presented.

In Chapter 3, the theory of HCRF is given: how the model is formulated, the conditional probabilistic model that uses for the classification, the way that learns its parameters and how inference can be achieved for a new observation sequence conditioned on these parameters.

In Chapter 4, our new model for action recognition, RI-HCRF, is first presented: the formulation of the model, the proposed algorithm for the automatic estimation of optimal number for hidden variables, the incorporation of the mixture of Student's *t* distribution to the model and finally the training and the inference of the model.

In Chapter 5, a detailed description of the datasets, the features and implementation is given. Moreover, the evaluation of the discriminatory properties of our model compared to three others, which where used as baseline, is described.

In Chapter 6, are summarized our contributions and discussion for future work.

Chapter 2

Related Work

Since the 1980s till nowadays, human action recognition remains a complex problem and a very active research area in computer vision. The information and multimedia explosion on the internet has resulted to a growing number of videos sequences and the imperative need for machine learning techniques to understand and analyze them. There are several surveys [7, 21–25] that provide a detailed overview of approaches reported in the literature, dedicated to human motion analysis and recognition from different research perspectives and communities.

According to Poppe's survey [7], the proposed recognition approaches for various human actions may be categorized into two main groups:

- classifiers used for direct classification of spatio-temporal features
- temporal state-space models used for action classification directly in the time domain.

The first group refers to classifiers that define the action without explicitly modeling variations in time, such as: kNN [2, 26–28], SVM [29–31], relevance vector machines (RVM) [32] and boosting frameworks [33–35]. These classifiers are usually used for approaches that are proposing new feature descriptors. Also, the aforementioned classifiers can be used either separately or combined together e.g., the bag-of-words framework [36] uses SVM or boosting for the classification of the frequency histogram of the feature descriptors). However, these nonstructural models assume that there is no correlation between descriptors, as they are considered to be identical and independently distributed (IID).

Apart from the conventional classifiers, probabilistic graphical models have more recently been applied. Their graph consist of nodes (states) and edges, so they are able to represent the probability distribution through their structure. These models belong to the second group of recognition approaches and can be either generative or discriminative. Generative models learn the joint distribution of observations and action class labels, while discriminative models learn the conditional distribution of action class labels given the observations.

A typical example of a generative approach is the Hidden Markov Model (HMM) [37], which has been traditionally, widely used in the literature. However, in order to keep the modeling of the joint distribution tractable, HMM makes two independence assumptions: 1) the probability of a certain hidden state at time n only depends on the hidden state at time n–1 and 2) the observations are considered independent of each other. Unlike the first assumption, the second one has a very limited validity and becomes a severe weakness of the HMM.

Discriminative models overcome the restrictive independence assumption of HMM, by allowing dependencies among the observations on different time scales. They can use multiple overlapping features, which is more suitable to complex data such as video sequencies. Also, discriminative models learn the differences between action categories instead of learning the modeling of each action separately. Two representative and widely used examples of the discriminative models are the Conditional Random Field (CRF) [18] and the Hidden Conditional Random Field (HCRF) [17].

CRF is a powerful classifier and commonly applied with good performance in activity recognition [38, 39]. Though, CRFs need a label to be assigned for each observation (e.g., each time step in a sequence), which is not provided in the majority of datasets and is a time-consuming procedure to be done manually. To overcome this limitation, HCRF was proposed as a method able to label sequences as a whole, by introducing an extra layer with hidden variables to the model's structure.

The HCRF model finds successful applications in various fields such as gesture recognition [40–42], action recognition [43–50], phone and speech recognition [51–55], handwriting recognition [56, 57], image and recordings segmentation [58,59], recognition of emotion and behavioral attributes [60–63], medicine [64,65], robotics [66–69] and many others. Aiming to improve the performance of the model, a large number of HCRF variants have been proposed by changing the topology, the training process and the feature function of the model. Song *et al.* [47] changed the topology of the model and proposed three multi-view HCRFs considering the action factorized into multiple views (e.g., body postures and hand shapes). Morency *et al.* [40] proposed a Latent-Dynamic Conditional Random Field (LDCRF), which includes a class label per observation and can be applied on unsegmented sequences. Vrigkas *et al.* [49] proposed active-HCRF+, a method that combines privileged information and active learning. The model needs to learn the weights of the privileged information that was added as an extra input of different modality. Wang *et al.* [44] introduced a Max-Margin HCRF, which maximizes the margin of hyperplanes between the correct and incorrect labels. Wang and Mori [43] proposed a model that combines large-scale global features and local patch features under the unified framework of HCRF. Also, Bousmalis *et al.* [62] proposed the infinite HCRF (iHCRF), a nonparametric model capable of automatically learning the optimal number of hidden states for a classification task. The model sets a hierarchical Dirichlet processes as prior to potentials of the model and learns its hyperparameters with an effective Markov-chain Monte Carlo sampling technique. Later, the same authors proposed a Variational HCRF [63], a generalized framework for infinite HCRF model and a novel variational inference approach that will converge faster reducing the computational cost. Finally, many hybrid models of HCRF have been proposed in the literature aiming to exploit the advantages of the combined methods in order to built a better classifier with a well-accentuated discrimanative ability [56,59,70,71].

Chapter 3

Hidden Conditional Random Fields

3.1 Introduction

The CRF is a graphical model widely used in various applications due to its efficient results. It combines interactions in consecutive labels and observed data. The data are not usually provided with their part labels that CRF needs and the manual assignment of the part labels will be a troublesome work especially for big data or video content.

To overcome these difficulties Hidden Conditional Random Fields (HCRFs) have been proposed as an extension of CRF. HCRF is a chain CRF introduced with an additional layer of structured hidden variables with dependencies among them. Hidden Conditional Random Fields are able to deal with more structured and complex data. Empirically, the classification performance in generative graphical models using hidden variables (e.g., HMM) has been improved and successfully addressed in a variety of problems. The integration of hidden variables also simplifies the complex joint distribution. The bag of words commonly use the conditional independence assumption which is relaxed in the HCRF case. The existence of direct link between the labels and the hidden states is also a useful structure. Finally, HCRFs are able to model spatial and temporal variations in observation sequences, which is a major capability for human activity that incorporate elementary or primary actions.

3.2 The Model Formulation

The classification task aims to map each observation sequence $x = \{x_1, x_2, ..., x_T\}$ to its actual label y. Every component x_j of the observation sequence is a local observation and is represented by a feature vector $\phi(x_j) \in \mathcal{R}^d$, where d is the dimensionality of the representation as it was defined at Quattoni *et al.* [17]. Every label y is a member of a set $\mathcal{Y}$ that denote the set of all possible actions, for example, $\mathcal{Y} = \{walk, run, jump\}$. For each time step t, the observation node is linked to the label node through the additional sub-structed layer of hidden variables. The hidden variables $h = \{h_1, h_2, ..., h_T\}$ are the part labels which are assigned to each observation during the training phase and are not observable in the training sequence. Hidden variables belong to a finite set $h_i \in \mathcal{H}$, where $\mathcal{H}$ represent the set of all possible hidden part labels. The use of hidden variables intents to capture the structure and movement patterns in the input space whilst they permit the inclusion of complex dependencies in the observation sequences.

An HCRF is expressed by an undirected graph. As a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, it consists of nodes $(\mathcal{V})$ that represent the variables and edges $(\mathcal{E})$ that express the correlation between them. The factor graph $\mathcal{G}$ is defined by a conditional probability distribution. Essentially, the graph $\mathcal{G}$ exploits the connectivity between hidden variables to discover potential dependencies. The illustration of the factor graph of the model is shown in Fig.3.1.

We assume that the graphical model forms a tree structure because, according to Quattoni *et al.* [17], the performance is not only equivalent to more densely connected graphical models but also reduces the computational complexity of the model. The rectangles in the graph correspond to unary potentials ψ_1, ψ_2 and pairwise potential function ψ_3. The linear combination of these potentials gives the potential function $\Psi(y, h, x; \theta) \in \mathcal{R}$, which is parametrised by θ. The model parameter θ consist of three components $\theta = \{\theta_1, \theta_2, \theta_3\}$ which are used to measure the compatibility between the observation sequences, the hidden variables and the class labels. Given the above

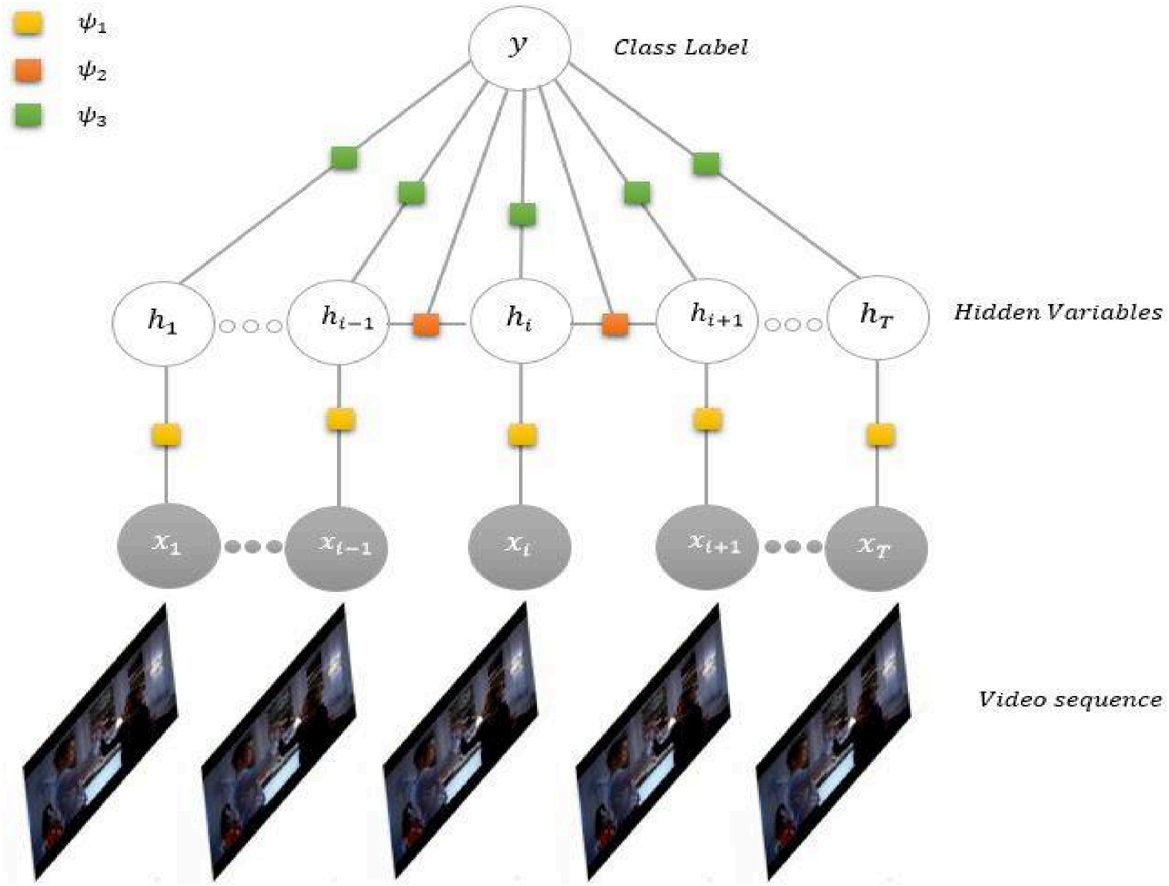


Figure 3.1: Illustration of the HCRF model. Each circle in the graph represents a variable, and each square represents a factor in the model.

definitions, the potential function is given by:

$$\Psi(y,h,x;\theta) = \sum_{j \in V} \theta_1 \cdot \psi_1(x_j,h_j) + \sum_{j \in V} \theta_2 \cdot \psi_2(y,h_j) + \sum_{(i,j) \in E} \theta_3 \cdot \psi_3(y,h_i,h_j). \tag{3.1}$$

The first two terms of the summation are called node terms and the last one edge term according to the nature of the relationship that they model.

The unary potential function $\theta_1 \cdot \psi_1(x_j,h_j)$ models the relationship between the hidden variable h_j and the feature vector x_j and is expressed by:

$$\theta_1 \cdot \psi_1(x_j,h_j) = \sum_{a \in H} \theta_{1_{aj}} \cdot \mathbb{1}_{\{h_j = a\}} \cdot x_j, \tag{3.2}$$

where $\mathbb{1}(\cdot)$ is the indicator function. This function is equal to 1 when its argument is true and equal to 0 otherwise. Obviously, the length of the vector θ_1 is $(d \times |\mathcal{H}|)$.

The unary potential function $\theta_2 \cdot \psi_2(y, h_j)$ models the relationship between the class label y and the hidden variable h_j and is expressed by:

$$\theta_2 \cdot \psi_2(y, h_j) = \sum_{a \in Y} \sum_{b \in H} \theta_{2_{a,b}} \cdot \mathbb{1}_{\{y=a\}} \cdot \mathbb{1}_{\{h_j=b\}}. \tag{3.3}$$

The length of the θ_2 is $(|\mathcal{Y}| \times |\mathcal{H}|)$.

The pairwise potential function $\theta_3 \cdot \psi_3(y, h_i, h_j)$ models the relationship between the class label y and the hidden variable h_j and h_j and is expressed by:

$$\theta_3 \cdot \psi_3(y, h_i, h_j) = \sum_{a \in Y} \sum_{b \in H} \sum_{c \in H} \theta_{3_{a,b,c}} \cdot \mathbb{1}_{\{y=a\}} \cdot \mathbb{1}_{\{h_i=b\}} \cdot \mathbb{1}_{\{h_j=c\}}. \tag{3.4}$$

This component (θ_3), represents the edges (links) between the hidden variables and it could be considered identical to the transition matrix of the HMM, though in HCRF model case there is a transition matrix for every possible class label. The length of the θ_3 is $(|\mathcal{Y}| \times |\mathcal{H}| \times |\mathcal{H}|)$.

3.3 Conditional Probabilistic Model

The conditional probabilistic model of the HCRF is defined by:

$$P(y, h|x, \theta) = \frac{\exp \Psi(y, h, x; \theta)}{\sum_{y' \in \mathcal{Y}} \sum_h \exp \Psi(y', h, x; \theta)}, \tag{3.5}$$

where y denotes the class labels, h denotes the hidden variables or the hidden part labels, x denotes the input observation sequence and θ denotes the parameters of the model. The denominator of the fraction is the partition function and it is a normalization constant term equivalent to the expectation of the unnormalised model over all possible classes y and all possible hidden variables h.

The posterior probability of a class label y , given an observation sequence x is calculated by marginalizing over all hidden variables h:

$$P(y|x, \theta) = \sum_h P(y, h|x, \theta) = \frac{\sum_h \exp \Psi(y, h, x; \theta)}{\sum_{y' \in \mathcal{Y}} \sum_h \exp \Psi(y', h, x; \theta)}. \tag{3.6}$$

If the hidden variables h are observed and there is a single class label y then the conditional probability of $P(h|X)$ becomes a regular CRF. By the use of the Bayes' rule in equations (3.5) and (3.6) we can compute the joint probability of assigning a set of hidden variables to an observation sequence when the class label and parameters are known:

$$P(h|y,x,\theta) = \frac{P(y,h|x,\theta)}{P(y|x,\theta)} = \frac{e^{\Psi(y,h,x;\theta)}}{\sum_h e^{\Psi(y,h,x;\theta)}}. \tag{3.7}$$

Similar to a CRF model, the aim is to maximize the conditional probability $P(y|x,\theta)$. The following objective function is used to learn the parameters θ:

$$\mathcal{L}(\theta) = \sum_i \log P(y_i|x_i;\theta) - \frac{1}{2\sigma^2}\|\theta\|^2, \tag{3.8}$$

where the first term the conditional log-likelihood on the input data and the second term is the log of a Gaussian prior, used as penalty term. The role of the penalty term is to avoid overfitting by assuming that the parameters θ of the model follow a Gaussian distribution with variance σ^2, $P(\theta) \sim \exp\big(-\frac{1}{2\sigma^2}\|\theta\|^2\big)$ to constrain $\|\theta\|$. The optimal weights $\theta^\star$ are learned by maximizing the objective function $\theta^\star = \arg\max_\theta \mathcal{L}(\theta)$.

3.4 Learning and Inference

The estimation of the optimal parameter $\theta^\star = \arg\max_\theta \mathcal{L}(\theta)$ can not be done analytically. Therefore, the iterative method of limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) [72] is employed. LBFGS is the most efficient and the most popular Quasi-Newton update formula among all gradient-based methods and works well for high-dimensional vectors. This iterative method computes repeatedly the objective function $\mathcal{L}(\theta)$ and its derivatives with respect to parameter θ. To reduce the computational complexity instead of storing and updating the entire inverse Hessian matrix, LBFGS stores and updates only the information from the past m iterations. The introduction of hidden states to the probabilistic model leads to a non convex objective function. As a result, the iterative method will usually get stuck in local extrema of the cost function, while the global extrema is not guaranteed that it will be reached.

The derivative of the objective function for a given sample of the training set may be written as:

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \theta} &= \sum_t \frac{\partial \mathcal{L}_t(\theta)}{\partial \theta} - \frac{\theta}{\sigma^2} \\
&= \sum_t \frac{\partial (\log(P(y_t|x_t;\theta))}{\partial \theta} - \frac{\theta}{\sigma^2} \\
&= \sum_t \frac{\partial \Big(\log \frac{\sum_h \exp \Psi(y_t,h,x_t;\theta)}{\sum_{y'\in\mathcal{Y}} \sum_h \exp \Psi(y',h,x_t;\theta)} \Big)}{\partial \theta} - \frac{\theta}{\sigma^2} \\
&= \sum_t \frac{\partial \big(\log \sum_h \exp \Psi(y_t,h,x_t;\theta) - \log \sum_{y'\in\mathcal{Y}} \sum_h \exp \Psi(y',h,x_t;\theta) \big)}{\partial \theta} - \frac{\theta}{\sigma^2}.
\end{aligned}
\tag{3.9}
$$

The estimation of the derivative of the objective function requires the calculation of the first order partial derivatives $\frac{\partial \mathcal{L}(\theta_k)}{\partial \theta_k}$ with respect to θ_k where $k \in \{1,2,3\}$. Therefore, the partial derivative of the the objective function takes the following form:

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_k} &= \sum_t \frac{\partial \big(\log \sum_h \exp \Psi(y_t,h,x_t;\theta) - \log \sum_{y'\in\mathcal{Y}} \sum_h \exp \Psi(y',h,x_t;\theta) \big)}{\partial \theta_k} - \frac{\theta_k}{\sigma^2} \\
&= \frac{\sum_h \exp \Psi(y_t,h,x_t;\theta) \cdot \frac{\partial \Psi(y_t,h,x_t;\theta)}{\theta_k}}{\sum_h \exp \Psi(y_t,h,x_t;\theta)} \\
&- \frac{\sum_{y'\in\mathcal{Y}} \sum_h \exp \Psi(y',h,x_t;\theta) \cdot \frac{\partial \Psi(y',h,x_t;\theta)}{\theta_k}}{\sum_{y'\in\mathcal{Y}} \sum_h \exp \Psi(y',h,x_t;\theta)} - \frac{\theta_k}{\sigma^2} \\
&= \sum_h P(h|y_t,x_t;\theta) \frac{\partial \Psi(y_t,h,x_t;\theta)}{\theta_k} - \sum_{y'\in\mathcal{Y}} \sum_h P(y',h|x_t;\theta) \frac{\partial \Psi(y',h,x_t;\theta)}{\theta_k} - \frac{\theta_k}{\sigma^2}
\end{aligned}
\tag{3.10}
$$

The derivative of the objective function is a time consuming task due to the number of hidden states. Assuming there are T observation sequences, the number of possible hidden variables h is $|\mathcal{H}|^T$. Equation (3.10) requires the calculation of two marginalized probabilities. These quantities can be estimated using the belief propagation algorithm [73] as follows:

$$
\forall y \in \mathcal{Y}, \quad Z(y|x;\theta) = \sum_h e^{\Psi(y,h,x;\theta)}, \tag{3.11}
$$

$$
\forall y \in \mathcal{Y},\ \forall j \in V,\ \forall \alpha \in \mathcal{H}, \quad P(h_j = \alpha | y, x; \theta) = \sum_{\mathbf{h}: h_j = \alpha} P(\mathbf{h}|y,x;\theta), \tag{3.12}
$$

$$\forall y \in \mathcal{Y},\ \forall (j,k) \in E,\ \forall \alpha \in \mathcal{H},\ \forall b \in \mathcal{H},\ P(h_j = \alpha, h_k = b | y, x; \theta) = \sum_{\mathbf{h}: h_j = \alpha, h_k = b} P(\mathbf{h}|y, x; \theta). \tag{3.13}$$

In equation (3.11), the partition function is described as the summation over all possible hidden variables. The marginal probability over an individual variable h_j is defined in (3.12) while the marginal probability over pairs of variables (h_j, h_k) is defined in (3.13). Thus, the first derivative with respect to θ_1, by taking into consideration (3.10) and (3.2) is:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1} &= \sum_h P(h|y_t, x_t; \theta) \frac{\partial \Psi(y_t, h, x_t; \theta)}{\theta_1} - \sum_{y' \in \mathcal{Y}} \sum_h P(y', h|x_t; \theta) \frac{\partial \Psi(y', h, x_t; \theta)}{\theta_1} - \frac{\theta_1}{\sigma^2} \\ &= \sum_h P(h|y_t, x_t; \theta) \sum_{j \in \mathcal{V}} \psi_1(x_t, h_j) - \sum_{y' \in \mathcal{Y}} \sum_h P(y', h|x_t; \theta) \sum_{j \in \mathcal{V}} \psi_1(x_t, h_j) - \frac{\theta_1}{\sigma^2}.\end{aligned}$$

By replacing (3.11) and (3.12) in the latter equation, it takes the form:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1} =& \sum_{j \in \mathcal{V}} \sum_{\alpha \in \mathcal{H}} P(h_j = \alpha | y_t, x_t; \theta) \psi_1(x_t, h_j) \\ &- \sum_{y' \in \mathcal{Y}} \sum_{j \in \mathcal{V}} \sum_{\alpha \in \mathcal{H}} P(h_j = \alpha, y' | x_t; \theta) \psi_1(x_t, h_j) - \frac{\theta_1}{\sigma^2}.\end{aligned} \tag{3.14}$$

In the same way, the first derivative with respect to θ_2 using (3.10) and (3.3) is:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta_2} &= \sum_h P(h|y_t, x_t; \theta) \frac{\partial \Psi(y_t, h, x_t; \theta)}{\theta_2} - \sum_{y' \in \mathcal{Y}} \sum_h P(y', h|x_t; \theta) \frac{\partial \Psi(y', h, x_t; \theta)}{\theta_2} - \frac{\theta_2}{\sigma^2} \\ &= \sum_h P(h|y_t, x_t; \theta) \sum_{j \in \mathcal{V}} \psi_2(y_t, h_j) - \sum_{y' \in \mathcal{Y}} \sum_h P(y', h|x_t; \theta) \sum_{j \in \mathcal{V}} \psi_2(y', h_j) - \frac{\theta_2}{\sigma^2}.\end{aligned}$$

By replacing (3.11) and (3.12) in the latter equation, it takes the form:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta_2} =& \sum_{j \in \mathcal{V}} \sum_{\alpha \in \mathcal{H}} P(h = \alpha | y_t, x_t; \theta) \psi_2(y_t, h_j) \\ &- \sum_{y' \in \mathcal{Y}} \sum_{j \in \mathcal{V}} \sum_{\alpha \in \mathcal{H}} P(h_j = \alpha, y' | x_t; \theta) \psi_2(y', h_j) - \frac{\theta_2}{\sigma^2}.\end{aligned} \tag{3.15}$$

Similarly, the first derivative with respect to θ_3 using (3.10) and (3.4) is:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta_3} &= \sum_{h \in \mathcal{H}} P(h|y_t, x_t; \theta) \frac{\partial \Psi(y_t, h, x_t; \theta)}{\theta_3} - \sum_{y' \in \mathcal{Y}} \sum_{h \in \mathcal{H}} P(y', h|x_t; \theta) \frac{\partial \Psi(y', h, x_t; \theta)}{\theta_3} - \frac{\theta_3}{\sigma^2} \\ &= \sum_{h \in \mathcal{H}} \sum_{h' \in \mathcal{H}} P(h|y_t, x_t; \theta) \sum_{(i,j) \in \mathcal{E}} \psi_3(y_t, h_i, h_j) \\ &- \sum_{y' \in \mathcal{Y}} \sum_{h \in \mathcal{H}} \sum_{h' \in \mathcal{H}} P(y', h|x_t; \theta) \sum_{(i,j) \in \mathcal{E}} \psi_3(y', h_i, h_j) - \frac{\theta_2}{\sigma^2}.\end{aligned}$$

By replacing (3.11) and (3.13) in the latter equation, it takes the form:

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta_3} &= \sum_{(i,j) \in \mathcal{E}} \sum_{\alpha \in \mathcal{H}} \sum_{b \in \mathcal{H}} P(h_i = \alpha, h_j = b|y_t, x_t; \theta) \psi_3(y_t, h_i, h_j) \\ &- \sum_{y' \in \mathcal{Y}} \sum_{(i,j) \in \mathcal{V}} \sum_{\alpha \in \mathcal{H}} \sum_{b \in \mathcal{H}} P(h_i = \alpha, h_j = b, y'|x_t; \theta) \psi_3(y', h_i, h_j) - \frac{\theta_2}{\sigma^2}\end{aligned} \tag{3.16}$$

Hence the value and the derivative of the objective function can be estimated and the model is able to learn its parameters $\theta^\star = \{\theta_1^\star, \theta_2^\star, \theta_3^\star\}$. Therefore, these parameters are used to find the class label for a given input observation, which constitutes the inference process. The class label for a new input is determined by:

$$y^\star = \arg\max_{y \in \mathcal{Y}} P(y|x; \theta^\star). \tag{3.17}$$

Also, in the case of linear HCRF models, the backward-forward inference algorithm can be used.

Chapter 4

Robust Incremental Hidden Conditional Random Fields

4.1 Introduction

HCRF seems to be a very promising approach in many application domains, due to its ability of relaxing strong independence assumption and exploiting temporal and spatial variations, via its graphical structure and links among the variables.

In the related literature, many researchers proposed a hybrid model. A combination of generative and discriminative models has been proved successfull in improving the performance of the classical models [70]. Motivated by this approach, some works combined HCRF with HMM. Soullard *et al.* [57] introduced a HMM-based weighting in the conditional probability of the HCRF which constrains the discriminative learning, yielding improved accuracy. On the other hand, Zhang *et al.* [45] used HMM to make hidden variables 'observable' to HCRF so the objective function can be convex.

However, the search of the optimal number of hidden variables remain a limitation of the model. The majority of the previous works, define the number of hidden variables in a intuitive manner or with exhaustive, computationally expensive and time consuming repeated evaluation of the model, for a given set of potential hidden variables. Bousmalis et al. [62] introduced Infinite Hidden Conditional Random Fields (iHCRF), a nonparametric model that estimates the number of hidden variables. The model assumes that the potentials of the HCRF are sampled directly from a set of Hierarchical Dirichlet Processeses and its hyperaparameters are learned using the sampling that removes hidden variables not presented in the samples.

In this work, a Robust Incremental Hidden Conditional Random Field (RI-HCRF) is proposed, which addresses two major issues in standard HCRFs. At first, the proposed model incrementally estimates the optimal number of hidden variables using a Viterbi-like procedure. Additionally, it uses a mixture of Student's *t* distribution as prior to the parameters of the model that leads to a model robust to outliers.

4.2 Formulation of the model

We consider a dataset $\mathcal{D} = \{x^{(k)}, y^{(k)}\}_{k=1}^{N}$ of N labeled observations. Here, the observation $x^k = \{x_1, x_2, ...x_T\}$ corresponds to the k^{th} video sequence that consists of T frames, and y^k is the k^{th} class label defined in a finite label set $\mathcal{Y}$. Each observation x^k can be represented by a feature vector $\phi(x_i^k) \in \mathcal{R}^d$, which is a collection of several features extracted from the i^{th} frame in a video sequence. Also, each observation sequence x^k is associated with a corresponding hidden variable sequence $h^k = \{h_1, h_2, ...h_T\}$, where $h^k \in \mathcal{H}$.

The model aims to find the most probable label y for a given observation x and the model parameter vector θ by maximizing the conditional probability $P(y|x;\theta)$. The highest conditional probability $P(y|x;\theta)$ indicates that the video sequence most likely belongs to class y. The conditional probability $P(y|x;\theta)$ is defined as the summation of exponentials of potential functions over all possible hidden variables h:

$$P(y|x,\theta) = \sum_h P(y,h|x,\theta) = \frac{\sum_h \exp \Psi(y,h,x;\theta)}{\sum_{y' \in \mathcal{Y}} \sum_h \exp \Psi(y',h,x;\theta)}, \tag{4.1}$$

where the potential function that specifies dependencies in the model, is given by:

$$\Psi(y,h,x;\theta) = \sum_{j\in V} \theta_1 \cdot \psi_1(x_j,h_j) + \sum_{j\in V} \theta_2 \cdot \psi_2(y,h_j) + \sum_{(i,j)\in E} \theta_3 \cdot \psi_3(y,h_i,h_j). \tag{4.2}$$

Following the work of HCRF presented in the previous Chapter, the RI-HCRF model can be represented by an undirected graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$. Each variable, observed (video frame) and unobserved (hidden variable) is a node $\mathcal{V}$ in the graphical model $\mathcal{G}$ and dependencies among them are presented by edges $\mathcal{E}$. The nodes are linked following the sequence of the video frames and form a linear chain graph.

4.3 Estimation of the number of hidden variables

In order to estimate the optimal number of hidden variables we propose a method which follows an iterative procedure. The method seeks for a sequence of hidden variables, where each variable participates in a maximum number of optimal paths. Therefore, variables with low participation in optimal paths are rejected.

We call the summation of unary potentials (Eqs. 3.2, 3.3) *node potentials* and the pairwise potential (Eq. 3.4) *edge potential*. The node potentials for a given label $y = \alpha$, for all observations and all possible hidden variables can be represented in a matrix form as follows:

$$Np = [np_{ij}]_{S\times T} = \begin{bmatrix} \theta_{1_{11}} \cdot x_1 + \theta_{2_{1\alpha}} & \theta_{1_{12}} \cdot x_2 + \theta_{2_{1\alpha}} & \dots & \theta_{1_{1T}} \cdot x_T + \theta_{2_{1\alpha}} \\ \theta_{1_{21}} \cdot x_1 + \theta_{2_{2a}} & \theta_{1_{22}} \cdot x_2 + \theta_{2_{2a}} & \dots & \theta_{1_{2T}} \cdot x_T + \theta_{2_{2\alpha}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1_{S1}} \cdot x_1 + \theta_{2_{S\alpha}} & \theta_{1_{S2}} \cdot x_2 + \theta_{2_{1\alpha}} & \dots & \theta_{1_{ST}} \cdot x_T + \theta_{2_{T\alpha}} \end{bmatrix},$$

where S is the number of hidden variables and T is the number of frames of the video sequence. The edge potentials for a given label $y = \alpha$, express the compatibility between a pair of hidden variables and can be represented by the following square matrix:

$$Ep = [ep_{ij}]_{S\times S} = \begin{bmatrix} \theta_{3_{11}} & \theta_{3_{12}} & \theta_{3_{13}} & \dots & \theta_{3_{1S}} \\ \theta_{3_{21}} & \theta_{3_{22}} & \theta_{3_{23}} & \dots & \theta_{3_{2S}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{3_{S1}} & \theta_{3_{S2}} & \theta_{3_{S3}} & \dots & \theta_{3_{SS}} \end{bmatrix}.$$

In order to bring all values of the Np matrix in the range [0,1] we use the normalizing function:

$$Np^{'} = \frac{1}{\max(Np) - \min(Np)} \cdot (Np - \min(Np)), \tag{4.3}$$

where $\min(Np)$ and $\max(Np)$ are the minimum and maximun value of Np respectively. This kind of normalization is also known as unity-based normalization. The normalization of the values of the Np matrix is done so that all the input variables have the same treatment in the model and its parameters are not scaled with respect to the units of the inputs. As a result, we will end up with smaller standard deviations, which can suppress the effect of outliers. Additionally, we construct a stochastic matrix based on the Ep, with each row summing to 1, by applying the following equation:

$$Ep' = \frac{Ep[i,j]}{\sum_{j=1}^{S} Ep[i,j]}. \tag{4.4}$$

In order to define the number of hidden variables, we employ multiple HMMs. A HMM is specified by:

- The set of hidden variables $h = \{h_1, h_2, ..., h_T\}$ and a set of parameters $\Lambda = \{\pi, A, B\}$.
    - π is a vector that collects the prior probabilities of $h_i,\ i \in \{1, 2, .., T\}$ being the first hidden variable of a state sequence.
    - A is a matrix that collects the transition probabilities of moving from one hidden variable to another.
    - B is the matrix that collects the emmision probabilities, which characterize the likelihood of a certain observation x, if the model is in hidden variable h_i.

Let us consider that the normalized node potentials are the entries of the emission probability matrix, the edge potentials are the entries of the transition probability matrix and there is a vector of initial probabilities π, where each hidden variable is equally probable to be first in the sequence:

$$\pi_{1\times\mathcal{S}} = \frac{1}{\mathcal{S}} \cdot 1_{\mathcal{S}}. \tag{4.5}$$

Given the above definitions we are able to determine an HMM and the optimal hidden variable sequence for a given label $y = \alpha$, using the Viterbi algorithm. This algorithm makes use of two variables: $\delta_t(i)$ and $\phi_t(i)$. The former variable estimates the probability of the most probable path ending in hidden variable i at time t while the latter keeps track of the "best path" ending in hidden variable i at time t:

$$\delta_t(i) = \max_{h_1,h_2,\ldots,h_{T-1}} P(h_1, h_2, ..., h_{T-1}, h_T = i, x_1, x_2, .., x_T | \pi, Ep, Np^{'}), \tag{4.6}$$

$$\phi_t(i) = \operatorname*{argmax}_{h_1,h_2,\ldots,h_{T-1}} P(h_1, h_2, ..., h_{T-1}, h_T = i, x_1, x_2, .., x_T | \pi, Ep, Np^{'}). \tag{4.7}$$

The illustration of the proposed method for the estimation of the optimal number of hidden variables is shown in Fig. 4.1

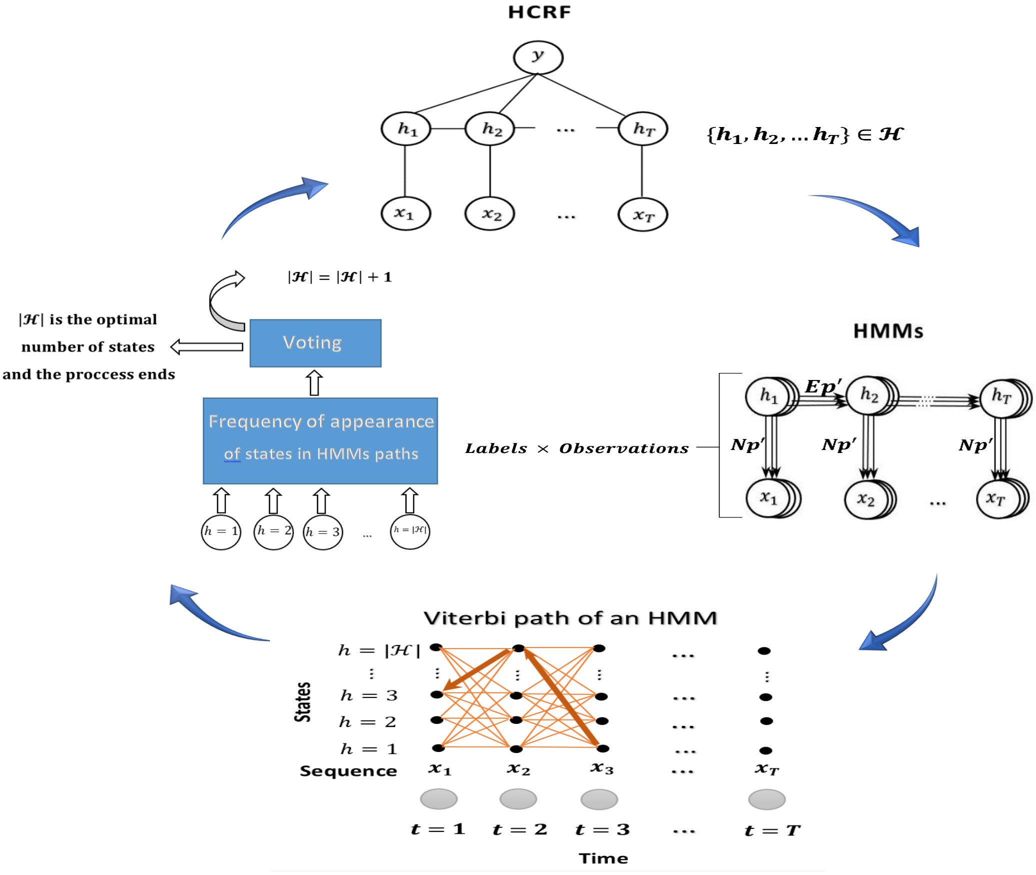


Figure 4.1: Illustration of the iterative and incremental method for the estimation of the optimal number of the hidden variables.

The proposed method learns the optimal number of hidden variables following an incremental learning approach. It starts by setting an initial number $S = |\mathcal{H}| \geq 1$ for the hidden variables and I for the maximum number of iterations. In each iteration, all optimal paths for every video sequence and for every label are estimated using the Viterbi algorithm. Also, the frequency of appearance of each hidden state in all paths is calculated. There is a termination criterion which is based on the frequency of each hidden variable to be lower than a threshold (τ). If this criterion is not satisfied, the number of hidden variables is increased by one and the process of calculating optimal paths and frequencies is repeated. If the criterion is satisfied, we move to the next iteration and a voting for the most probable number of hidden variables in the current iteration takes place. When all iterations are finished the optimal number of hidden variables is the one with the majority of votes. A detailed description of the proposed method is presented in Algorithm 4.1.

Algorithm 4.1 Estimation of the optimal number of hidden states $S^\star$

```
Require: $\mathcal{D} = \{x^{(k)}, y^{(k)}\}_{k=1}^{N}, S, I$
Ensure: $S^\star$
for $iteration = 1$ to $I$ do
    while 1 do
        Set $S$ the initial number of hidden variables and initialize $\theta$ randomly
        for each $x^{(k)} = \{x_1, x_2, \ldots, x_T\} \in \mathcal{D},\ \alpha \in \mathcal{Y}$, do
            $Np[m,n] = \sum_{m\in\mathcal{H}}(\theta_{1_{mn}} \cdot 1_{\{h_n=m\}} x_n^{(k)}) + \sum_{\alpha\in\mathcal{Y}} \sum_{m\in\mathcal{H}}(\theta_{2_{m\alpha}} \cdot 1_{\{h_n=m\}} \cdot 1_{\{y=\alpha\}})$
            $Np' = f(Np)$ {the normalized node potentials matrix}
            $Ep[m,n] = \sum_{a\in Y} \sum_{m\in H} \sum_{n\in H} \theta_{3_a,m,n} \cdot 1_{\{y=a\}} \cdot 1_{\{h_b=m\}} \cdot 1_{\{h_c=n\}}$
            $Ep' = Ep[m,n] = \frac{Ep[m,n]}{\sum_{n=1}^{S} Ep[m,n]}$ {the edge potentials matrix}
            for $i = 1$ to $S$ do
                $\delta_1(i) = \pi_i Np'[i,1]$
                $\phi_1(i) = 0$
            end for
            for $t = 2$ to $T$ do
                for $i = 1$ to $S$ do
                    $\delta_t(i) = Np'[i,t] \cdot \max_i(\delta_{t-1}(i) Ep'[i,t])$
                    $\phi_t(i) = \operatorname{argmax}_i(\delta_{t-1}(i) Ep'[i,t])$
                end for
            end for
            $h_T = \operatorname{argmax}_i(\delta_T(i))$
            for $t = T - 1$ to 1 do
                $h_t = \phi_{t+1}(h_{t+1})$
            end for {optimal path for the $k_{th}$ observation and the label $\alpha$ is $\{h_1, h_2, \ldots, h_T\}$}
        end for
        if frequency $f(\cdot)$ of each hidden variable $\geq$ threshold then
            $S = S + 1$
        else
            vote as $S_{temp}(iteration) = S$
            break
        end if
    end while
end for
$S^\star = \max(f(S_{temp}))$ is the number with the majority of votes
```

4.4 A student's t-mixture prior on the model parameters

A finite mixture model is a statistical distribution, derived by the combination or the mixture of other distributions. Each individual distribution, part of the model, is called mixture component while the weights associated with each component are called mixture weights. This approach is powerful and flexible in modeling complex data since it exploits and combine the properties of each individual probability density function. Over the years, finite mixture models have been applied in many fields due to their nature of being efficient and mathematically tractable. Great attention has been paid to Gaussian mixture models [74–76] because of their computational convenience. However, Gaussian mixture models have shown sensitivity not only to outliers but also to small amount of data. As a result, Student's t mixture models have been proposed to overcome the problems of Gaussian models.

The Student's t mixture model has a pdf with heavier or longer tails and gives reduced weight to the observations that are in the tail area, in the estimation of its parameters. In that way, it provides robustness to outliers and less extreme estimates of the posterior probabilities of component membership of the mixture model, according to McLachlan and Peel [77]. Additionally, each Student's t component originates from a wider class of elliptically symmetric distributions with an additional robustness tuning parameter called the degrees of freedom ν.

Let us assume that the parameters θ of the RI-HCRF follow a mixture model with three Student's t components instead of following a Gaussian distribution $P(\theta) \sim \exp\left(-\frac{1}{2\sigma^2}\|\theta\|^2\right)$. Taking into consideration that the parameter vector θ describes three different relationships among observations, hidden variables and labels, we expect that each component will correspond and to one of these relationships. In that way, the method relies on the partitioning of the parameter vector using a Student's t mixture model and identify, preserve and enhance the differences between these partitions for a better classification result. The choice of Student's t components has been made because of their robustness.

However, by making the above assumption the problem of setting the mixture weights arises. In order to find the best weights for the mixture model in advance, we will need to do an exhaustive search and check multiple combinations of probable values of mixture weights. To avoid that computational prohibitive approach, we decide, at the end of each iteration of parameters training, to estimate dynamically the best fitted mixture model for parameters θ.

Let vector parameter $\theta = \{\theta'_1, \theta'_2, \ldots, \theta'_M\}$ consists of M weights, where $M = (d \times |\mathcal{H}|) + (|\mathcal{Y}| \times |\mathcal{H}|) + (|\mathcal{Y}| \times |\mathcal{H}| \times |\mathcal{H}|)$. Each parameter θ follows a univariate t-distribution with mean μ, variance σ^2 and $\nu \in [0, \infty)$ degrees of freedom when, given the weight u, the parameter θ has the univariate normal with mean μ and variance σ^2/u:

$$\theta|\nu, \sigma^2, \mu, u \sim N(\mu, \sigma^2/u), \tag{4.8}$$

and weight u follows a Gamma distribution parameterized by ν:

$$u \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2}). \tag{4.9}$$

By integrating out the weights from the joint density leads to the density function of the marginal distribution:

$$\begin{aligned} p(\theta; \nu, \mu, \lambda) &= \int_0^\infty N(\theta|\mu, u/\sigma^2)\Gamma\big(u|\frac{\nu}{2}, \frac{\nu}{2}\big)du \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\Big(\frac{\lambda}{\pi\nu}\Big)^{\frac{1}{2}}\Big(1 + \frac{\lambda(\theta-\mu)^2}{\nu}\Big)^{-\frac{\nu+1}{2}}, \end{aligned} \tag{4.10}$$

where the inverse scaling parameter λ (similar to precision) is the reciprocal of variance ($\lambda = (\sigma^2)^{-1}$). Also, it can be shown that for $\nu \to \infty$ the Student's t-distribution tends to a Gaussian distribution. Moreover, for $\nu > 1$, μ is the mean of θ and for $\nu > 2$, the variance of θ is $\nu(\lambda(\nu-2))^{-1}$. The t-distribution is used to estimate probabilities based on incomplete data or small samples.

The best fitted mixture with Student's t-components can be obtained by maximizing the likelihood function using the EM algorithm [77]. A K-component mixture of t-distributions is given by:

$$\phi(\theta, \Omega) = \sum_{i=1}^{K} \pi_i p(\theta; \nu_i, \mu_i, \lambda_i), \tag{4.11}$$

where θ denotes the observed data vector and $\Omega = \{\Omega_i\}_{i=1}^K$ the mixture parameter set with $\Omega_i = \{\pi_i, \nu_i, \mu_i, \lambda_i\}$. For the mixing proportions of the i^{th} component π_i, we have that:

$$0 \leq \pi_i \leq 1, \quad i \in \{1, 2, \ldots, K\}, \ \sum_{i=1}^{K} \pi_i = 1. \tag{4.12}$$

The complete-data vector for the EM framework is given by:

$$\theta_c = (\theta^T, z_1^T, z_2^T, \ldots, z_M^T, u_1^T, u_2^T, \ldots, u_3^T)^T, \tag{4.13}$$

where $z_1, \ldots, z_M$ are the component-label vectors and $z_{ij} = (z_j)_i$ defines whether the observation $\theta'_j, j \in \{1, 2, \ldots, M\}$ belongs to the i^{th} component or not, by taking the value 1 and 0 respectively. Despite the augmentation of the observed data with z_j, it is convenient to view the observed data as still being incomplete and introduce into the complete-data vector the additional missing data $u_1, \ldots, u_M$ because the variances depend on the degrees of freedom. So, the E-step on the $(t+1)^{th}$ iteration of the EM algorithm requires the calculation of the posterior probability that the datum θ'_j belongs to the i^{th} component of the mixture:

$$z_{ij}^{t+1} = \frac{\pi_i^t p(\theta'_j; \nu_i^t, \mu_i^t, \lambda_i^t)}{\sum_{m=1}^{K} p(\theta'_j; \nu_m^t, \mu_m^t, \lambda_m^t)}, \tag{4.14}$$

and the expectation of the weights for each observation:

$$u_{ij}^{t+1} = \frac{\nu_i^t + 1}{\nu_i^t + \lambda_i(\theta'_j - \mu_i)^2}. \tag{4.15}$$

The update equations of the respective mixture model parameters are provided by maximizing the log-likelihood of the complete data:

$$\pi_i^{t+1} = \frac{1}{M} \sum_{j=1}^{M} z_{ij}^t, \tag{4.16}$$

$$\mu_i^{t+1} = \frac{\sum_{j=1}^{M} z_{ij}^t u_{ij}^t \theta'_j}{\sum_{j=1}^{M} z_{ij}^t u_{ij}^t}, \tag{4.17}$$

$$\lambda_i^{t+1} = \frac{\sum_{j=1}^{M} z_{ij}^t u_{ij}^t (\theta'_j - \mu_i^t)^2}{\sum_{j=1}^{M} z_{ij}^{t+1}}. \tag{4.18}$$

At last, the degrees of freedom for each component are computed as the solution to the equation:

$$\log\Big(\frac{\nu_i^{t+1}}{2}\Big) - \psi\Big(\frac{\nu_i^{t+1}}{2}\Big) + 1 - \log\Big(\frac{\nu_i^t + 1}{2}\Big)$$
$$+\frac{\sum_{j=1}^{M} z_{ij}^t (\log u_{ij}^t - u_{ij}^t)}{\sum_{j=1}^{M} z_{ij}^t} + \psi\Big(\frac{\nu_i^t + 1}{2}\Big) = 0, \tag{4.19}$$

where $\psi(x) = \frac{\partial(\ln(\Gamma(x))}{\partial(x)}$ is the digamma function. There is no closed form solution for ν, so it has to be found numerically. An extended description for the EM algorithm for Student's *t* mixture model can be found in [77].

4.5 Learning and Inference

Aiming to maximize the conditional probability $P(y|x,\theta)$, RI-HCRF use the following objective function to train the parameters θ:

$$\mathcal{L}(\theta) = \underbrace{\sum_{i=1}^{N} \log P(y_i|x_i;\theta)}_{\mathcal{L}_1} + \underbrace{\log\Big(\sum_{k=1}^{K} \pi_k p(\theta;\nu_k,\mu_k,\lambda_k)\Big)}_{\mathcal{L}_2}, \tag{4.20}$$

where the first term $\mathcal{L}_1$ is the conditional log-likelihood of the input data and the second term $\mathcal{L}_2$ represents the best fitted Student's *t* mixture model on parameter vector θ, obtained by the EM algorithm described in 4.4. The optimal weights $\theta^\star$ are learned by maximizing the objective function $\theta^\star = \underset{\theta}{\text{argmax}}\mathcal{L}(\theta)$.

The derivative of the objective function for a given sample of the training set may be written as:

$$\frac{\partial\mathcal{L}(\theta)}{\partial\theta} = \frac{\partial\mathcal{L}_{1_t}(\theta)}{\partial\theta} + \frac{\partial\mathcal{L}_{2_t}(\theta)}{\partial\theta}. \tag{4.21}$$

In section 3.4 a detailed description for the calculation of the derivative of the conditional log-likelihood $\mathcal{L}_{1_t}$ was presented. The analytical expression of $\mathcal{L}_{2_t}$ is given by:

$$
\begin{aligned}
\mathcal{L}_{2_t} &= \log\Big(\sum_{k=1}^{K}\pi_k p(\theta;\nu_k,\mu_k,\lambda_k)\Big)\\
&= \log\Big(\pi_1 p(\theta;\nu_1,\mu_1,\lambda_1)+\ldots+\pi_k p(\theta;\nu_k,\mu_k,\lambda_k)\Big)\\
&= log\Bigg(\pi_1\frac{\Gamma(\frac{\nu_1+1}{2})}{\Gamma(\frac{\nu_1}{2})}\Big(\frac{\lambda_1}{\pi\nu_1}\Big)^{\frac{1}{2}}\Big(1+\frac{\lambda_1(\theta-\mu_1)^2}{\nu_1}\Big)^{-\frac{\nu_1+1}{2}}+\ldots\\
&+\pi_k\frac{\Gamma(\frac{\nu_k+1}{2})}{\Gamma(\frac{\nu_k}{2})}\Big(\frac{\lambda_k}{\pi\nu_k}\Big)^{\frac{1}{2}}\Big(1+\frac{\lambda_k(\theta-\mu_k)^2}{\nu_k}\Big)^{-\frac{\nu_k+1}{2}}\Bigg).
\end{aligned}
\tag{4.22}
$$

Therefore, the derivative of $\mathcal{L}_{2_t}$ can be written as:

$$
\begin{aligned}
\frac{\partial\mathcal{L}_{2_t}(\theta)}{\partial\theta} &= \frac{\partial\log\Big(\pi_1 p(\theta;\nu_1,\mu_1,\lambda_1)+\ldots+\pi_k p(\theta;\nu_k,\mu_k,\lambda_k)\Big)}{\partial\theta}\\
&= \frac{\frac{\partial(\pi_1 p(\theta;\nu_1,\mu_1,\lambda_1))}{\partial\theta}+\ldots+\frac{\partial(\pi_k p(\theta;\nu_k,\mu_k,\lambda_k))}{\partial\theta}}{\pi_1 p(\theta;\nu_1,\mu_1,\lambda_1)+\ldots+\pi_k p(\theta;\nu_k,\mu_k,\lambda_k)}\\
&= \frac{\sum_{k=1}^{K}\pi_k p(\theta;\nu_k,\mu_k,\lambda_k)\Big(-\frac{\lambda_k(\nu_k+1)(\theta-\mu_k)}{1+\lambda_k(\theta-\mu_k)^2}\Big)}{\sum_{k=1}^{K}\pi_k p(\theta;\nu_k,\mu_k,\lambda_k)}
\end{aligned}
\tag{4.23}
$$

The training task for the learning of $\theta^\star$ can be done using the LBFGS method [72], since the value and the derivative of the objective function may be calculated. The inference task is to find the label for a given input using these parameters:

$$
y^\star = \underset{y\in\mathcal{Y}}{\text{argmax}} P(y|x;\theta^\star).
$$

A description of the overall RI-HCRF method is presented in Algorithm 4.2.

Algorithm 4.2 RI-HCRF Training

```
Require: $\mathcal{D} = \{x^{(k)}, y^{(k)}\}_{k=1}^{N}, S_0 = |\mathcal{H}|$, where $\{h_1, h_2, ...h_T\} \in \mathcal{H}$ , $I$, MaxIter, $\tau$
Ensure: $\theta^\star$
 1: while $iteration < I$ do
 2:    Initialize $\theta$ randomly
 3:    for each $x^{(k)} = \{x_1, x_2, \ldots, x_T\} \in \mathcal{D},\ \alpha \in \mathcal{Y}$, do
 4:       Create an HMM with parameters according to Eqs (4.3,4.4,4.5)
 5:       Calculate the Viterbi path for each HMM
 6:    end for
 7:    Calculate the $f(\cdot)$, frequency of appearance, for
       $h_i,\ i = \{1, 2, ...T\}$ in all paths                                  Alg. 4.1
 8:    if the $f(h_i) \geq \tau$ then
 9:       $S = S + 1$, go to 2.
10:    else
11:       vote $S_{temp}(iteration) = S,\ iteration = iteration + 1$,
          $S = S_0$, go to 1.
12:    end if
13: end while
14: Given $S^\star = \max(f(S_{temp}))$, initialize $\theta$ randomly
15: while iteration < MaxIter do
16:    Estimate the parameters of the mixture of Students t for $\theta$ using the EM algorithm
17:    Update RI-HCRF gradients (Eq. (4.21)) using LBFGS
18: end while
19: return $\theta^\star$
```

Chapter 5

Experimental Results

In this chapter we evaluate the classification performance of our proposed method RI-HCRF and the performance of three other methods used as baseline: SVM [78]), CRF [18] and HCRF [17]. The models will be evaluated on six publicly available benchmark datasets and aim to tackle different problems of action recognition. The structure of this chapter is organized as follows: first, a description of the datasets will be given; second, the feature representation that is used in each dataset will be described; next, implementation details will be given and finally, a performance analysis of the conducted experiments will be made.

5.1 Description of datasets

Arm Gesture Dataset: a gesture database that was used by Wang *et al.* [1] and consist of 724 sequences and six gestures: Expand Horizontally (EH), Expand Vertically (EV), Shrink Vertically (SV), Point and Back (PB), Double Back (DB) and Flip Back (FB).

In EH gesture, the actor begins with both arms close to the hips, moves both arms laterally apart and retracts back to the resting position. In EV gesture, the arms move vertically apart and return to the resting position. In SV gesture, both arms begin from the hips, move vertically together and back to the hips. In PB gesture, the actor points with one hand and beckons with the other. In DB gesture, both arms beckon towards the user. Finally, in FB gesture, the actor simulates holding a book with one hand while the other hand makes a flipping motion. The gestures were performed by 13 actors, and an average of 90 gestures were collected per class. The illustration of these gestures are shown in Fig 5.1, where the green arrows are the motion trajectory of the fingertip and their direction symbolizes the direction of the performed gesture.

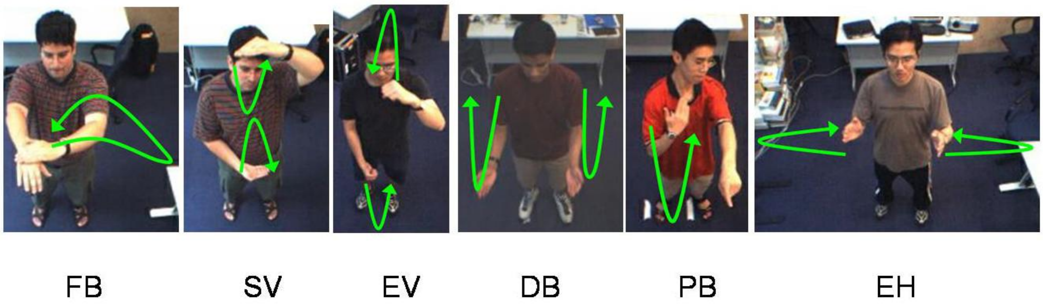


Figure 5.1: Representative frames of the Arm Gesture Dataset [1].

Weizmann dataset: a dataset that was used by Blank *et al.* [2] and consists of 10 different actions: bend, jack for jumping jack, jump for jump forward on two legs, pjump for jump in place on two legs, run, side for gallop sideways, skip, walk, wave1 for wave one hand and the last wave2 for wave two hands. This dataset contains 93 videos at 180×144 pixel resolution, 25 frames per second and records of 9 different actors performing the actions in front of a static camera and homogeneous outdoor backgrounds. Lena, of the actors, performs the actions run, skip and walk twice, starting the action from the opposite direction each time. It also provides a background subtraction mask for each video frame. Figure 5.2 illustrates a sample frame of each action from the Weizmann dataset.

Parliament dataset: a dataset that was created by Vrigkas *et al.* [3] and contains 228 video sequences at 320 × 240 pixel resolution and 25 frames per second. This dataset keeps records of 20 different individuals speaking in the Greek parliament and includes 3 statements of human behavior: friendly, aggressive and neutral, depending on the intensity of the political speech and the specific individual's movements (Fig. 5.3).This dataset helps to understand human interactions.

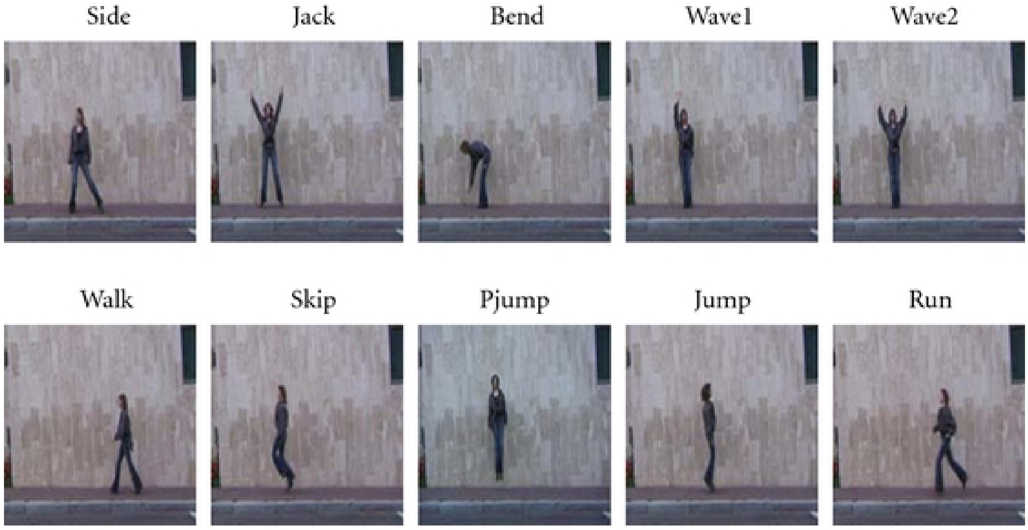


Figure 5.2: Representative frames of the Weizmann dataset [2].

Figure 5.3: Representative frames of the Parliament dataset [3].

Two-person interaction (TPI): a dataset that was introduced by Yun *et al.* [4] and includes approximately 300 video sequences, which represent the way that two-person interact. This interaction is captured by a Microsoft Kinect sensor. Seven different actors perform 8 different interactions: approaching, departing, kicking, pushing, shaking hands, hugging, exchanging objects, and punching (Fig. 5.4). The dataset also contains three-dimensional coordinates of 15 joints for each person at each frame.

TV human interaction (TVHI) a dataset proposed by Patron *et al.* [5] with 300 video sequences, range from 30 to 600 frames, collected from over 20 different TV shows. In these videos are represented four kinds of interactions:hand shakes, high fives, hugs and kisses. Each class contains 50 video sequences while the remaining 100 videos belong to negative examples (e.g., clips that do not contain any of the aforementioned interactions). The intra degree and inter-class diversity among the videos is very high due to the different actors, variations in scale and angle of the

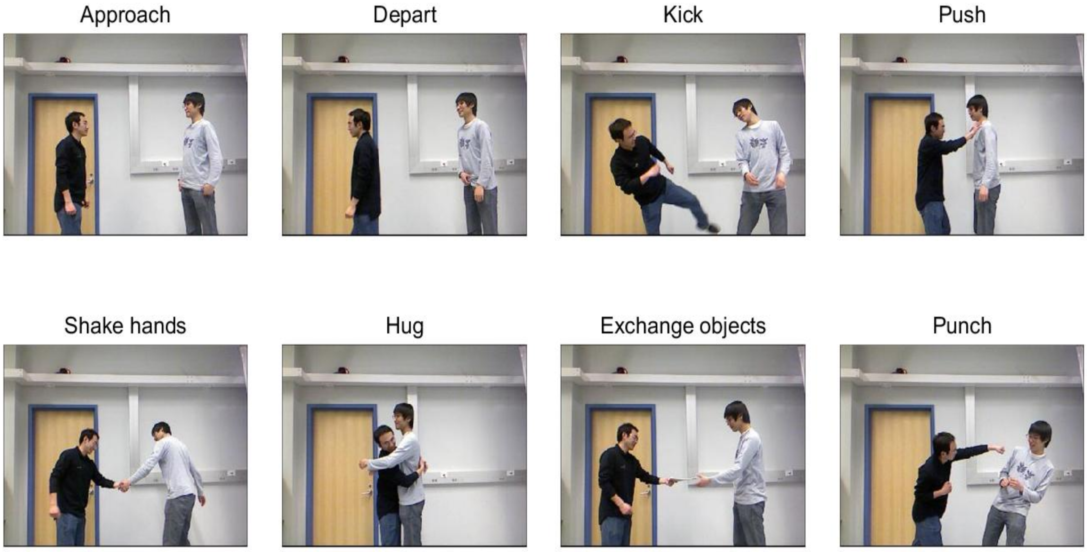


Figure 5.4: Representative frames of the TPI dataset [4].

camera and make this dataset suitable example of real world problems. Figure 5.5 depicts some representative frames of the TVHI dataset.

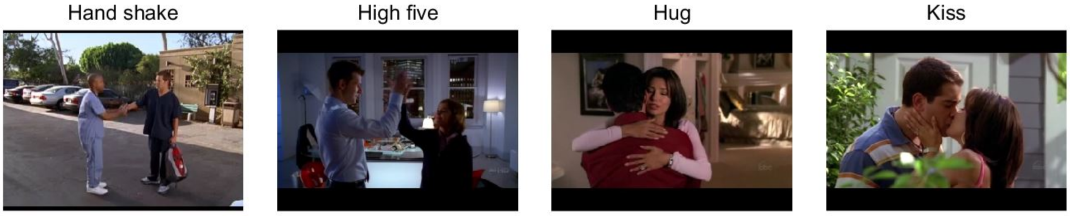


Figure 5.5: Representative frames of the TVHI dataset [5].

Unstructured social activity attribute (USAA): The USAA dataset was used by Fu *et al.* [6] and it contains eight different semantic class videos of social occasions:birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance, and wedding reception (Fig. 5.6). It has a collection of around 100 videos per class for training and testing. Each video is annotated with 69 attributes that can be broken down into five broad classes: actions, objects, scenes, sounds, and camera movement. The USAA dataset is a subset of Columbia consumer video database (CCV), which contains 9,317 YouTube videos and 20 semantic categories.

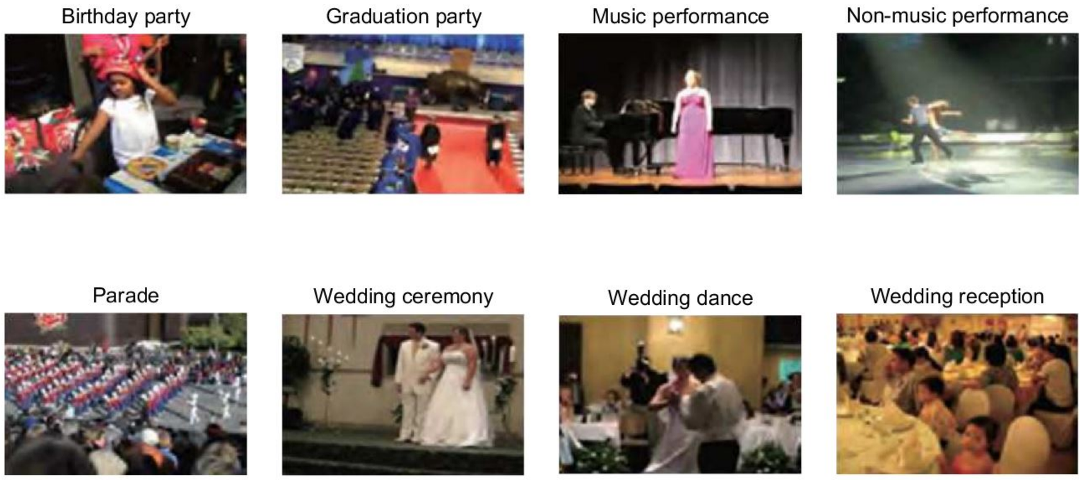


Figure 5.6: Representative frames of the USAA dataset [6].

5.2 Features

Body joint configuration: For the evaluation of our method on Arm Gesture dataset we used the features that were provided by the dataset. Each actor was asked to perform these gestures in front of a stereo camera and a stereo-tracking algorithm [79] was used to estimate the head, torso, arms and forearms, by a 3D cylindrical body model. A redundant parameterization composed of 2D joint angles and 3D Euclidean coordinates for left/right shoulders and elbows define the 20D input observations, corresponding to each frame.

Spatio-temporal interest points (STIP): We used STIP [11] as a compact representation of the video sequences of Weizmann dataset. The descriptors are Histograms of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF), computed on a 3D video patch in the neighborhood of each detected STIP, which captures the human motion between frames. The patch is partitioned into a grid with 3x3x2 spatio-temporal blocks. HOG and HOF descriptors are then computed for all blocks and are concatenated into a 72-dimensional vector and 90-dimensional vector descriptors respectively.

Synchronized feature cues: In order to provide more information to the classifier and enchance the efficiency of STIP features we used synchronized feature cues proposed by Vrigkas *et al.* [49, 50]. A fusion of audio-visual features is used in the Parliament dataset where the visual features are represented by STIP. Similarly, in TVHI the visual features are represented by STIP and the provided annotations of the dataset (locations and bounding boxes for the actors, head orientations and the pairs of intracting subjects). Mel-frequency cepstral coefficients (MFCCs) [80] and their first and second order derivatives form the audio feature vector of dimension 39. The audio signal was processed over 10 ms using a Hamming window with 25% overlap and sampled at 16 KHz. Finally, the TPI dataset, was represented by pose-visual feature vectors. Specifically, the pose refers not only to the positions of the locations of the joints for each person in each frame, but also to six more feature types concerning joint distance, joint motion, plane, normal plane, velocity, and normal velocity as described by Yun *et al.* [4]. The visual representation of the video data, is summarized by STIP features.

Features produced by 3D Convolutional Neural Networks (C3D): A method used by Du Tran *et al.* [81], that extracts features by convolving 3D kernels with the cube formed by stacking multiple adjacent frames. As a result the method is able to select features from both spatial and temporal dimensions and consequently it is able to capture motion information encoded in multiple consecutive frames. Given the input frames, C3D produces multiple channels. The combination of information from all channels is the final feature representation by a 1024D vector for Parliament, TPI and USAA datasets and by a 256D vector for TVHI dataset.

All types of features used for each dataset in our method are summarized in Table 5.1.

5.3 Models and implementation details

The evaluation of our model has been conducted on gesture, action and behavior datasets, which were mentioned above. Also, we trained a multi-class Support Vector Machine (SVM) [82], using (LIBSVM [78]), as a first baseline framework. SVM builds a model that assigns new examples to one class or the other by maximizing the margin of the hyper-planes that is used to separate the observation data. In this model we set the one against all multi-class method, the cost $c = 1$ and the width of the Gaussian kernel $g = 0.2$. The values for these parameters were determined by cross-validation by combining the values $c = 10^{k_1}$, with $k_1 \in \{-1, 0, 1, 10, 100\}$ and $gamma = k_2 \cdot 10^{-1}$, with $k_2 \in \{1, 2, \ldots, 10, 50, 100\}$.

As a second baseline, we trained a single CRF chain model [18] where every class had a corresponding hidden variable. In the standard CRF the model predicts a class

Datasets	Features	Dimension
Arm Gesture [1]	Angles and coordinates of joints	20
Weizmann [2]	STIP	162
Parliament [3]	STIP	162
	MFCC	39
	C3D	1024
TPI [4]	STIP	162
	Pose	15
	C3D	1024
TVHI [5]	STIP	162
	Head orientations	2
	MFCC	39
	C3D	256
USAA [6]	C3D	1024

Table 5.1: Types and dimension of features for each dataset.

label for each frame in a sequence, in contrast to the rest of the models which predict a label for the entire sequence.

A hidden conditional random field model, was also used as a third baseline method, to prove the efficiency of the HCRF model to learn the hidden dynamics between the video sequences of different action categories.

Finally, in RI-HCRF the threshold used in the proposed method for the automatic learning of the optimal number of hidden variables was set to take values from a discrete set $\tau \in \{0.001, 0.005, 0.01, 0.02, 0.05\}$ and the number of iteration was set to $I = 30$. The number of components for the mixture of Student's t distribution was set to K=3. The model parameters were randomly initialized and the experiments were repeated 5 times.

The number of hidden variables varied from 3 to 18 for both HCRF and RI-HCRF models, in order to examine the performance of the models with the same number of hidden variables.

5.4 Results and Discussion

5.4.1 Evaluation on the Arm Gesture dataset

In the Arm Gesture dataset we used 10-fold cross validation to split into training and test sets. In Fig.5.7 the classification accuracy with respect to the number of hidden variables for both HCRF and RI-HCRF models is depicted.

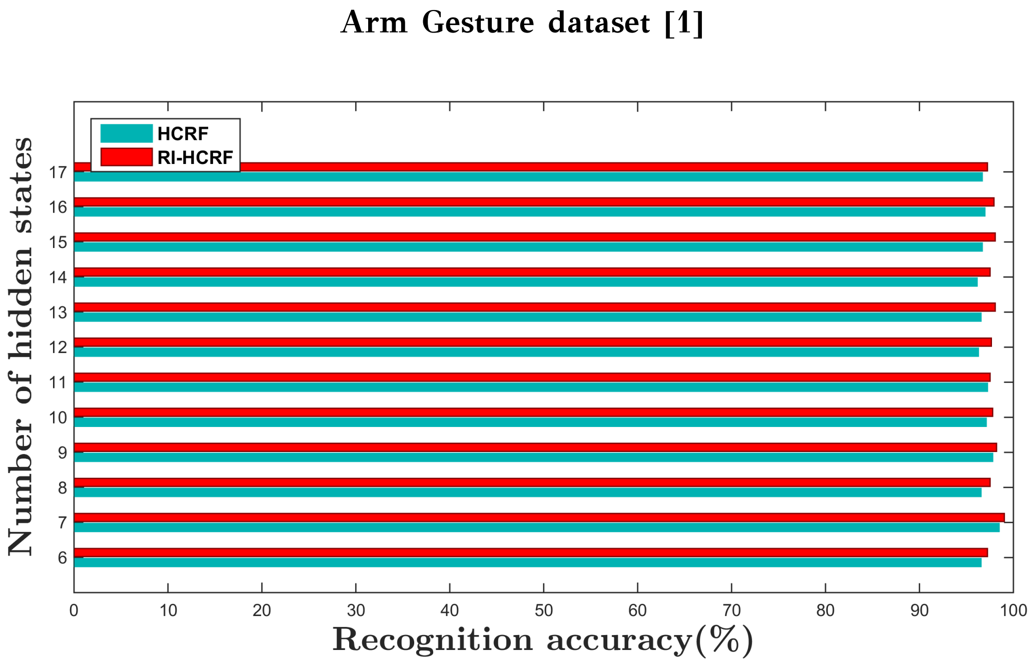


Figure 5.7: Classification accuracy with respect to the number of hidden states for the Arm Gesture dataset [1].

We can observe that the proposed model (RI-HCRF) seems to perform better than the standard HCRF. Also, the recognition accuracy for both models appears to decrease as the number of hidden variable increases. In the majority of hidden states, they have a similar behavior or similar fluctuations of accuracy. In particular, when the accuracy for one method increases/decreases from state to state the same applies for the other method.

Fig.5.8 indicates the optimal number of hidden states that was predicted by implementing Algorithm 4.1. The prediction of the optimal number of hidden states is accurate, since it is identical to the state that both models achieved the highest recognition score in the conducted experiments (Fig.5.7).

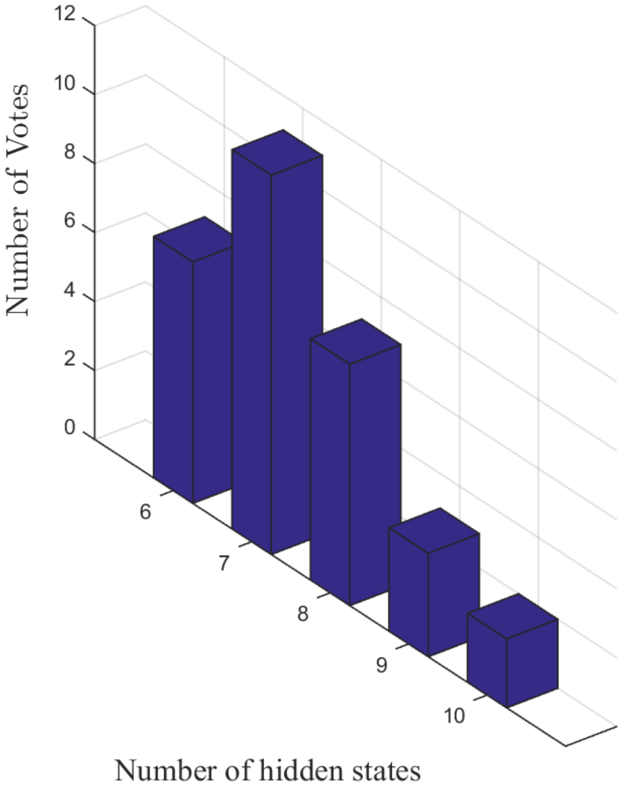
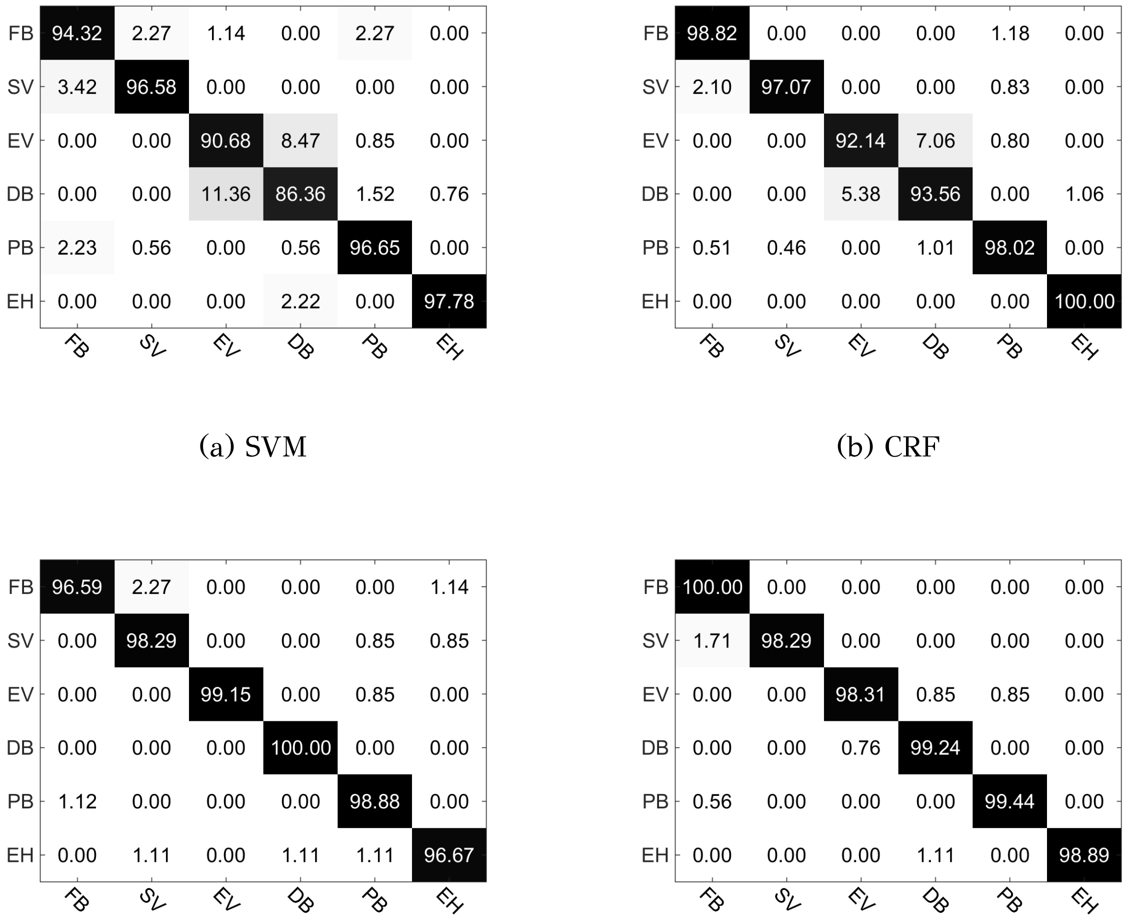


Figure 5.8: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the Arm Gesture dataset [1]. The number of hidden states that do not appear in the horizontal axis received zero votes.

The resulting confusion matrices of the 3 baseline methods and the RI-HCRF are presented in Fig. 5.9. The matrices for HCRF and RI-HCRF correspond to the optimal number of hidden variables. It is worth mentioning that the SVM and CRF incorrectly classify the Expand Vertically (EV) and Double Back (DB) gestures. Moreover, CRF, HCRF and RI-HCRF perfectly recognize Expand Horizontally (EH), Double Back (DB) and Flip Back (FB) respectively.

A detailed representation of the results for the Arm Gesture dataset is shown in Table 5.2. The table presents the mean recognition accuracy and its standard deviation. Comparing the results of the proposed RI-HCRF with the other three baseline methods it may be seen that RI-HCRF achieves the highest accuracy score with the lowest standard deviation.

Arm Gesture dataset [1]



	FB	SV	EV	DB	PB	EH
FB	94.32	2.27	1.14	0.00	2.27	0.00
SV	3.42	96.58	0.00	0.00	0.00	0.00
EV	0.00	0.00	90.68	8.47	0.85	0.00
DB	0.00	0.00	11.36	86.36	1.52	0.76
PB	2.23	0.56	0.00	0.56	96.65	0.00
EH	0.00	0.00	0.00	2.22	0.00	97.78

(a) SVM

	FB	SV	EV	DB	PB	EH
FB	98.82	0.00	0.00	0.00	1.18	0.00
SV	2.10	97.07	0.00	0.00	0.83	0.00
EV	0.00	0.00	92.14	7.06	0.80	0.00
DB	0.00	0.00	5.38	93.56	0.00	1.06
PB	0.51	0.46	0.00	1.01	98.02	0.00
EH	0.00	0.00	0.00	0.00	0.00	100.00

(b) CRF

	FB	SV	EV	DB	PB	EH
FB	96.59	2.27	0.00	0.00	0.00	1.14
SV	0.00	98.29	0.00	0.00	0.85	0.85
EV	0.00	0.00	99.15	0.00	0.85	0.00
DB	0.00	0.00	0.00	100.00	0.00	0.00
PB	1.12	0.00	0.00	0.00	98.88	0.00
EH	0.00	1.11	0.00	1.11	1.11	96.67

(c) HCRF

	FB	SV	EV	DB	PB	EH
FB	100.00	0.00	0.00	0.00	0.00	0.00
SV	1.71	98.29	0.00	0.00	0.00	0.00
EV	0.00	0.00	98.31	0.85	0.85	0.00
DB	0.00	0.00	0.76	99.24	0.00	0.00
PB	0.56	0.00	0.00	0.00	99.44	0.00
EH	0.00	0.00	0.00	1.11	0.00	98.89

(d) RI-HCRF

Figure 5.9: Confusion matrices for the classification results for the Arm Gesture dataset [1].

		Categories					
Method	**Overall**	**FB**	**SV**	**EV**	**DB**	**PB**	**EH**
SVM	93.64 ± 3.6	94.32 ± 6.1	96.58 ± 4.4	90.68 ± 12.0	86.36 ± 6.1	96.65 ± 5.3	97.78 ± 4.6
CRF	96.53 ± 1.4	98.82 ± 4.0	97.07 ± 4.9	92.14 ± 9.4	93.56 ± 6.2	98.02 ± 3.1	100 ± 0
HCRF	98.48 ± 0.7	96.59 ± 5.6	98.29 ± 3.6	99.15 ± 2.6	100 ± 0	98.88 ± 2.3	96.67 ± 7.5
RI-HCRF	99.03 ± 1.1	100 ± 0	98.29 ± 3.5	98.31 ± 3.6	99.24 ± 2.4	99.44 ± 1.7	98.89 ± 3.5

Table 5.2: Averaged recognition accuracies of all methods for the Arm Gesture dataset [1] (mean ± st. dev.).

5.4.2 Evaluation on the Weizmann dataset

A 9-fold leave-one-actor-out cross validation was used in order to split the Weizmann dataset into training and test sets. The results of the classification accuracy with respect to the number of hidden variables for both HCRF and RI-HCRF models are presented in Fig.5.10.

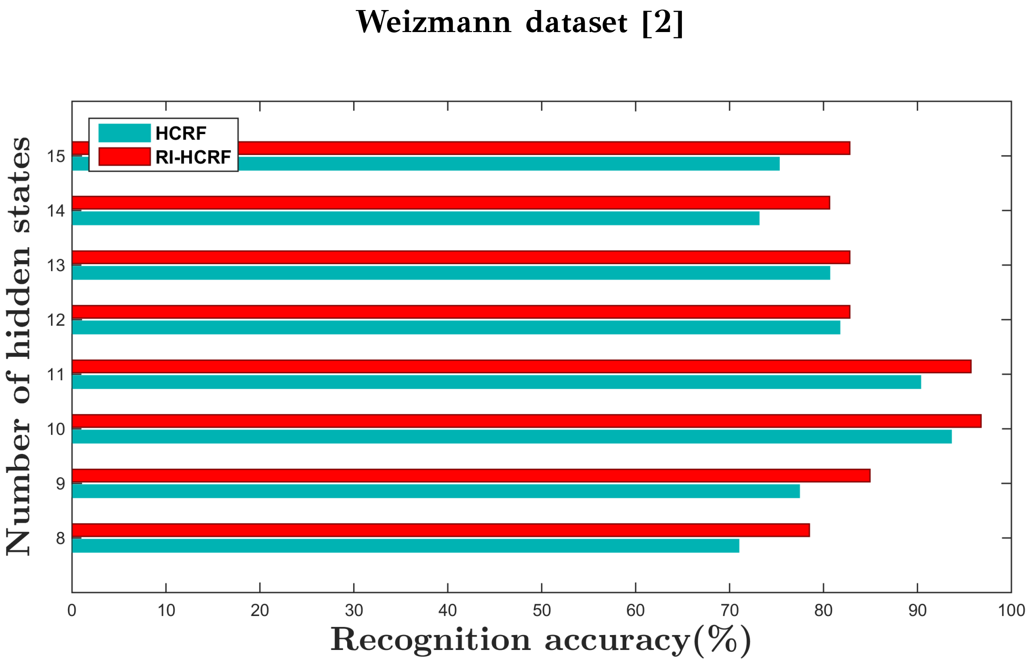


Figure 5.10: Classification accuracy with respect to the number of hidden states for the Weizman dataset [2].

It may be observed that RI-HCRF outperforms the standard HCRF, since it has always better accuracy score for every tested hidden state. HCRF reach 93.59% accuracy for 10 hidden states and with the same number of hidden states RI-HCRF reach 96.67%. Similar to Arm Gesture dataset, these two models seems to share the same pattern of accuracy variations.

The results of the automatic estimation of the optimal number of hidden variables are presented in Fig.5.11, where the value of 10 for the hidden states has collected the 73% of the votes. The predicted optimal number of hidden variables confirms the experimental results obtained by exhaustive search shown in Fig.5.10.

The confusion matrices for all models are depicted in Fig. 5.12. RI-HCRF classifies correctly seven of the ten actions, in contrast to other baseline methods, which reach 100% accuracy for 5 or less actions. Moreover, the smallest classification error between classes belongs to the proposed RI-HCRF method. It is worth to be noted that the

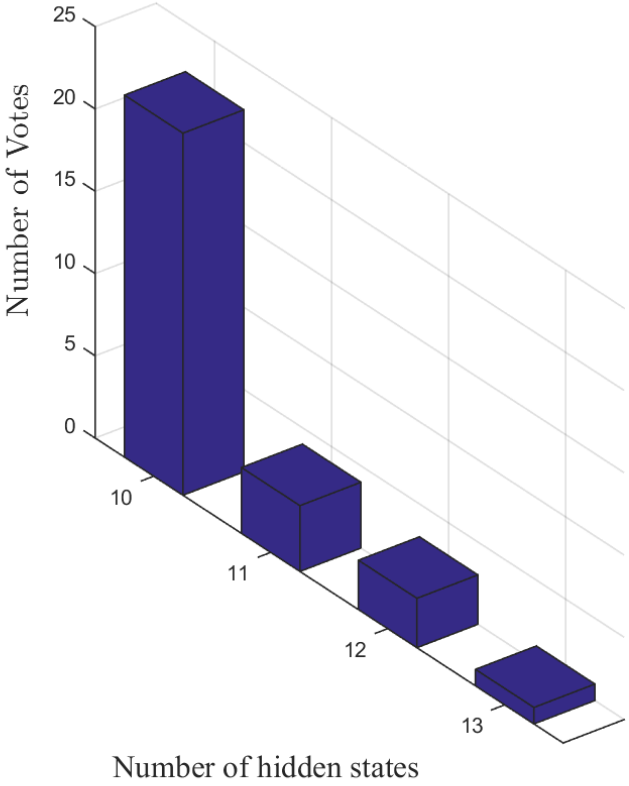


Figure 5.11: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the Weizmann dataset [2]. The number of hidden states that do not appear in the horizontal axis received zero votes.

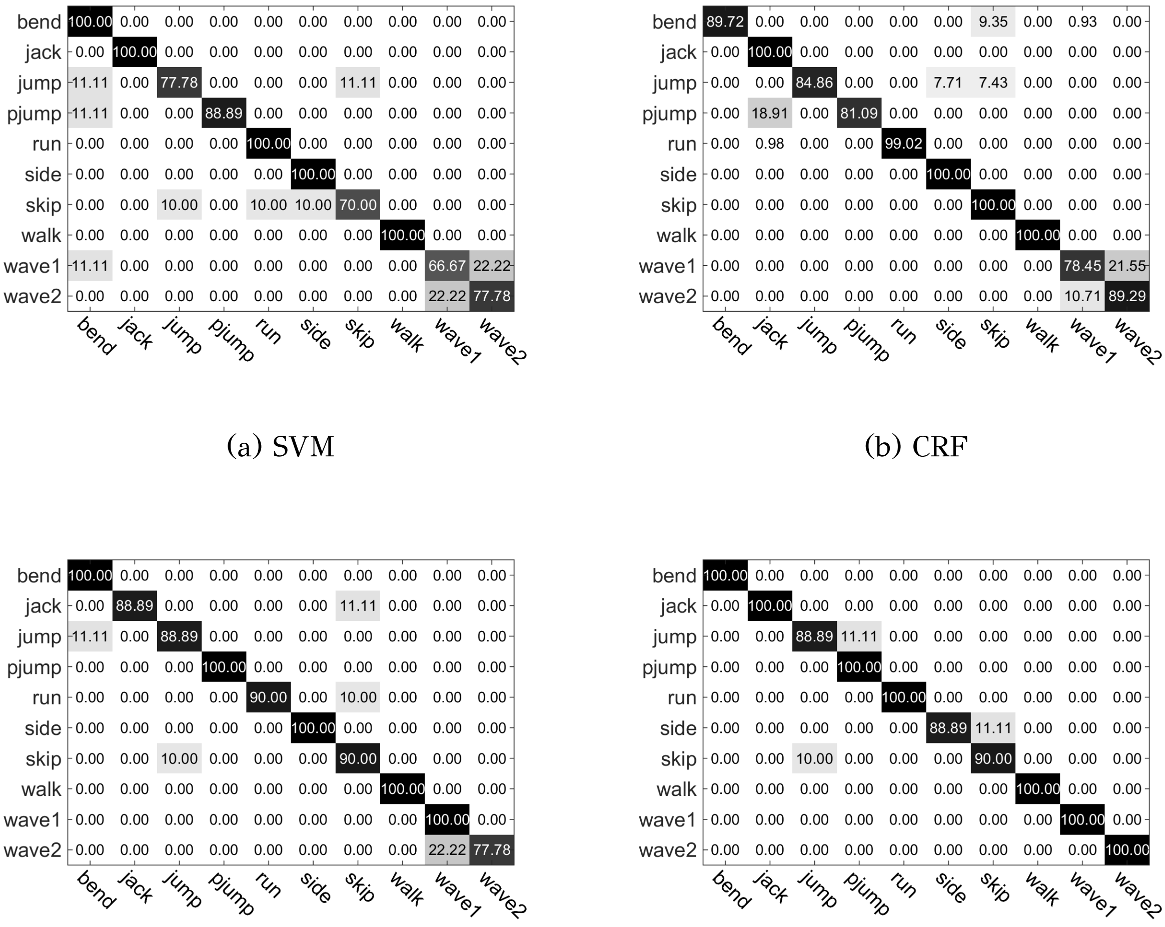
Weizmann dataset has actions that are very similar to each other. So, the large intra-class variability implies that different actions may be strongly confused.

The average recognition accuracy and its standard deviation for each model and each action category are presented in Table 5.3. The large values of standard deviation are a result of the leave one actor out cross validation that was used to split this dataset. There was one video sequence for testing of each action class, apart from Lena's case.

		Categories									
Method	**Overall**	**Bend**	**Jack**	**Jump**	**Pjump**	**Run**	**Side**	**Skip**	**Walk**	**Wave1**	**Wave2**
SVM	88.18 ± 13.9	100 ± 0	100 ± 0	77.78 ± 44.1	88.89 ± 33.3	100 ± 0	100 ± 0	77.00 ± 44.1	100 ± 0	66.67 ± 50.0	77.78 ± 44.1
CRF	95.48 ± 4.0	89.72 ± 14.38	100 ± 0	84.86 ± 34.1	81.09 ± 36.3	99.02 ± 2.6	100 ± 0	100 ± 0	100 ± 0	78.45 ± 50.0	89.29 ± 33.3
HCRF	93.59 ± 4.8	100 ± 0	88.89 ± 33.3	88.89 ± 33.3	100 ± 0	90 ± 16.67	100 ± 0	90.0 ± 33.3	100 ± 0	100 ± 0	77.78 ± 44.1
RI-HCRF	96.67 ± 5.0	100 ± 0	100 ± 0	88.89 ± 33.3	100 ± 0	100 ± 0	88.89 ± 33.3	90 ± 33.3	100 ± 0	100 ± 0	100 ± 0

Table 5.3: Averaged recognition accuracies of all methods for the Weizmann dataset [2] (mean ± st. dev.).

Weizmann dataset [2]



	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jack	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jump	11.11	0.00	77.78	0.00	0.00	0.00	11.11	0.00	0.00	0.00
pjump	11.11	0.00	0.00	88.89	0.00	0.00	0.00	0.00	0.00	0.00
run	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
side	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
skip	0.00	0.00	10.00	0.00	10.00	10.00	70.00	0.00	0.00	0.00
walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
wave1	11.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.67	22.22
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.22	77.78

(a) SVM

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	89.72	0.00	0.00	0.00	0.00	0.00	9.35	0.00	0.93	0.00
jack	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jump	0.00	0.00	84.86	0.00	0.00	7.71	7.43	0.00	0.00	0.00
pjump	0.00	18.91	0.00	81.09	0.00	0.00	0.00	0.00	0.00	0.00
run	0.00	0.98	0.00	0.00	99.02	0.00	0.00	0.00	0.00	0.00
side	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
skip	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00
walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	78.45	21.55
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.71	89.29

(b) CRF

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jack	0.00	88.89	0.00	0.00	0.00	0.00	11.11	0.00	0.00	0.00
jump	11.11	0.00	88.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pjump	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
run	0.00	0.00	0.00	0.00	90.00	0.00	10.00	0.00	0.00	0.00
side	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00
skip	0.00	0.00	10.00	0.00	0.00	0.00	90.00	0.00	0.00	0.00
walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.22	77.78

(c) HCRF

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jack	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jump	0.00	0.00	88.89	11.11	0.00	0.00	0.00	0.00	0.00	0.00
pjump	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
run	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00
side	0.00	0.00	0.00	0.00	0.00	88.89	11.11	0.00	0.00	0.00
skip	0.00	0.00	10.00	0.00	0.00	0.00	90.00	0.00	0.00	0.00
walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(d) RI-HCRF

Figure 5.12: Confusion matrices for the classification results for the Weizmann dataset [2].

5.4.3 Evaluation on the Parliament dataset

The Parliament dataset will be described in two parts, with respect to the the feature representation that was used. The first part will refer to the results extracted from dataset's representation of features which were produced by a 3D convolutional neural network. As for the second part, it will refer to the results extracted from a representation of fusion and synchronized features. In both parts, a 5-fold cross validation was used to create the train and test sets.

Parliament dataset expressed by C3D features

The recognition accuracy of HCRF and RI-HCRF models as the number of hidden variables increases is given in Fig. 5.13. We may observe that steep variations occur

Parliament dataset [3]

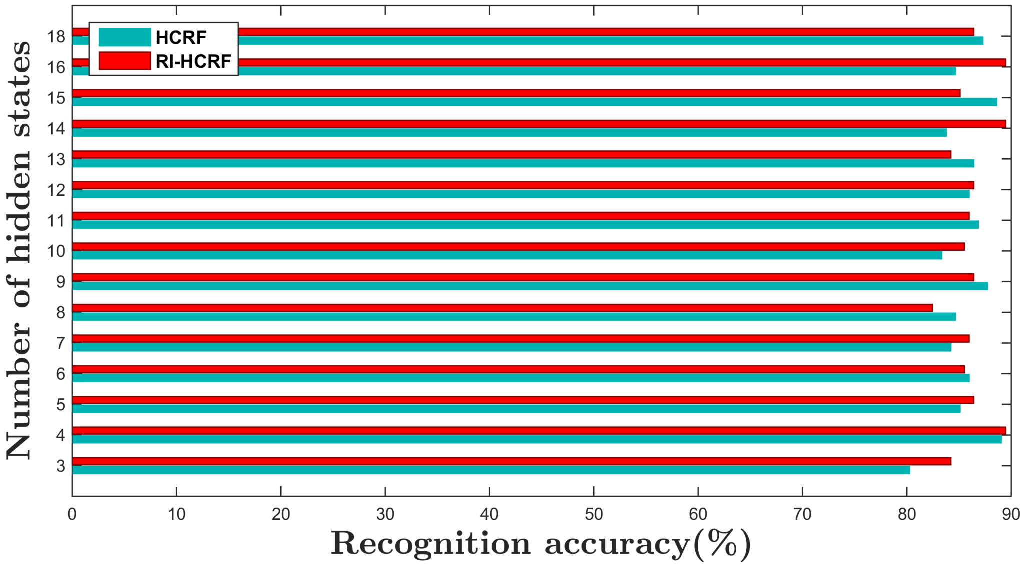


Figure 5.13: Classification accuracy with respect to the number of hidden states for the Parliament dataset [3] using C3D features.

in accuracy score for both models. In brief, the recognition accuracy falls and rises drastically from one hidden variable to another. These variations may be caused by the initialization of the parameters of the models. The highest accuracy score of HCRF (89.04%) is reached for 4 hidden states and the highest accuracy score of RI-HCRF (89.46%) is reached for 4, 14 and 16 hidden states. In this dataset as the number of hidden variables increases the accuracy score is kept high for both models, in contrast to Arm Gesture and Weizmann datasets behavior.

The automatic estimation of the optimal number of hidden variables indicated number 4 to be the most suitable candidate (Fig. 5.14) and concurrent agreed with the experimental results presented in Fig. 5.13.

The confusion matrices of all methods indicate a high confusion between aggressive and friendly behavior, while neutral behavior seems to be more recognizable as it is shown in Fig. 5.15

The average recognition accuracy and its standard deviation for each model and each action category are presented in Table 5.4. RI-HCRF has the lowest deviation in the overall accuracy rate and in the majority of action categories. Comparing the mean accuracy rate, our proposed method outperforms the state of the art models that where used as baseline in this work.

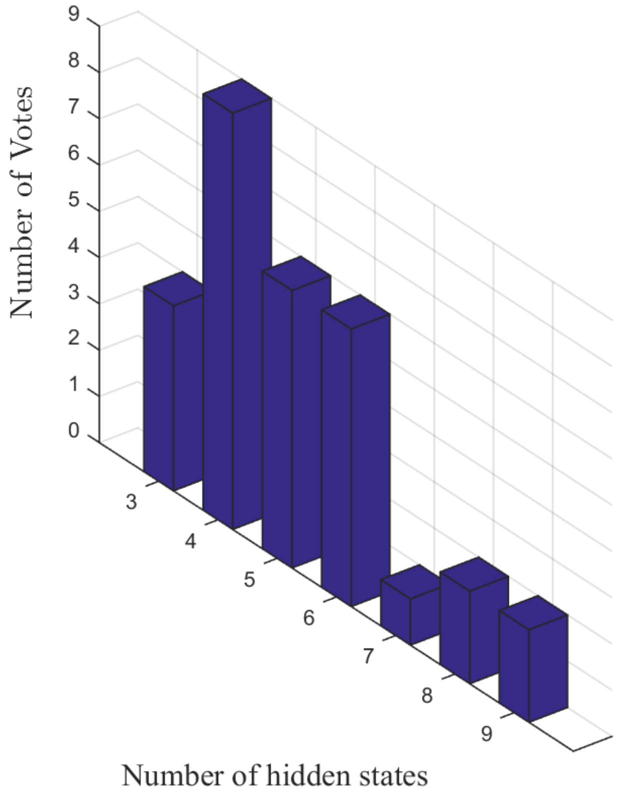


Figure 5.14: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the Parliament dataset [3] using C3D features. The number of hidden states that do not appear in the horizontal axis received zero votes.

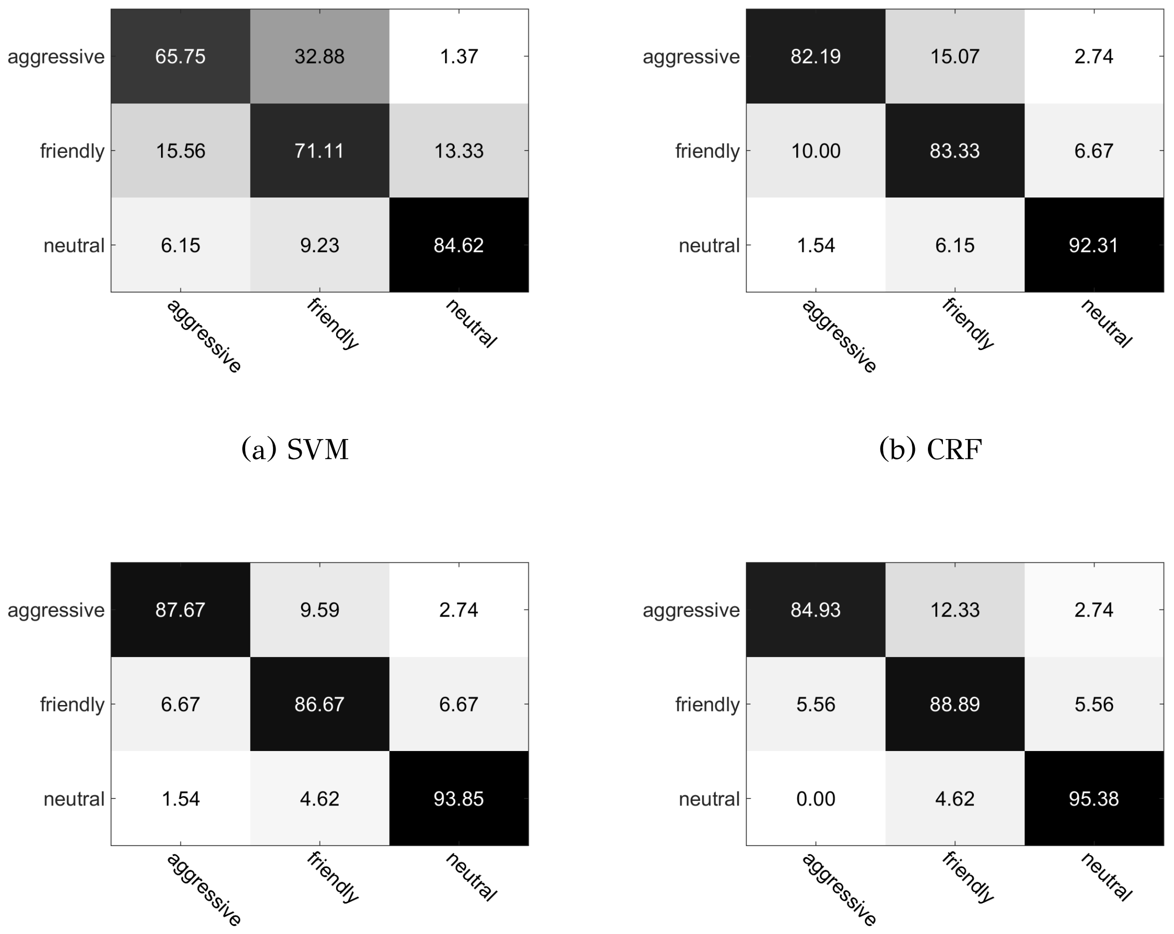
		Categories		
Method	**Overall**	**Aggressive**	**Friendly**	**Neutral**
SVM	73.28 ± 9.0	65.75 ± 8.7	71.11 ± 13.2	84.62 ± 14.4
CRF	85.54 ± 7.0	82.19 ± 12.8	83.33 ± 15.2	92.31 ± 5.4
HCRF	89.04 ± 8.1	87.67 ± 10.6	86.67 ± 12.7	93.85 ± 6.4
RI-HCRF	89.46 ± 6.5	84.93 ± 12.8	88.89 ± 11.7	95.38 ± 4.2

Table 5.4: Averaged recognition accuracies of all methods for the Parliament dataset [3] using C3D features (mean ± st. dev.).

Parliament dataset expressed by synchronized and fusion features

The graphical representation of the recognition accuracy rate with respect to changes in the number of hidden variables, for the proposed method RI-HCRF and

Parliament dataset [3]



(a) SVM

(b) CRF

(c) HCRF

(d) RI-HCRF

Figure 5.15: Confusion matrices for the classification results for the Parliament dataset [3] using C3D features.

the standard HCRF, is presented in Fig. 5.16. The RI-HCRF and HCRF methods require four hidden states to reach the best recognition accuracy, which is 82.85% and 85.5% respectively. As the number of hidden states increases, the accuracy for both models, is getting higher scores but not equivalent to the highest. Also, the accuracy for both models is the same or approximately equal when 14, 15, 16 and 17 hidden variables are used.

The results of automatic estimation of the optimal number of hidden variables are shown in Fig. 5.17 and lead to the conclusion that the use of 4 hidden states will be enough and able to exploit and distinguish sub-stuctures in the models. The predicted optimal number of hidden variables is equal to the optimal number derived

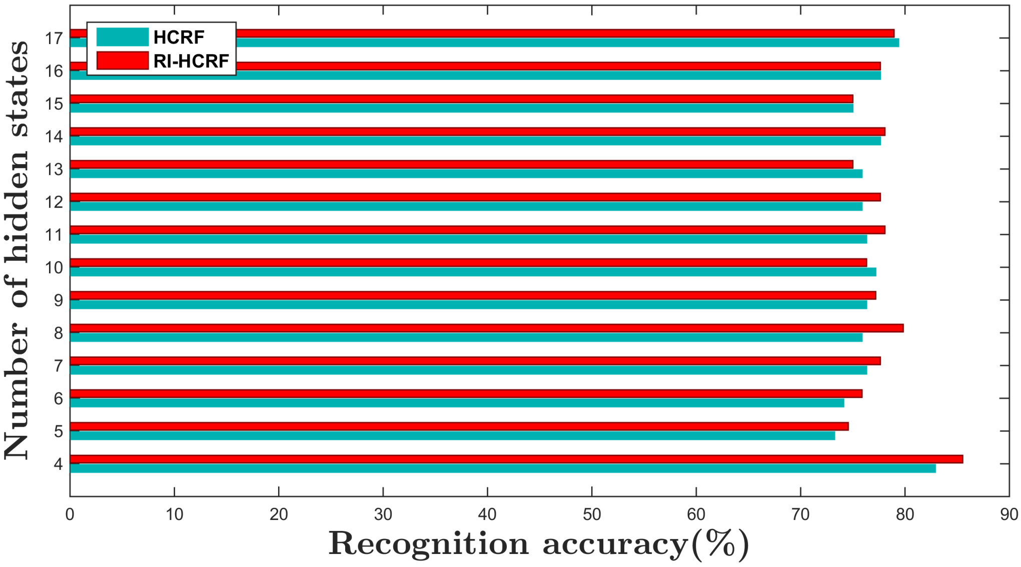


Figure 5.16: Classification accuracy with respect to the number of hidden states for the Parliament dataset [3] using synchronized and fusion features cues.

from the exhaustive search (Fig. 5.16).

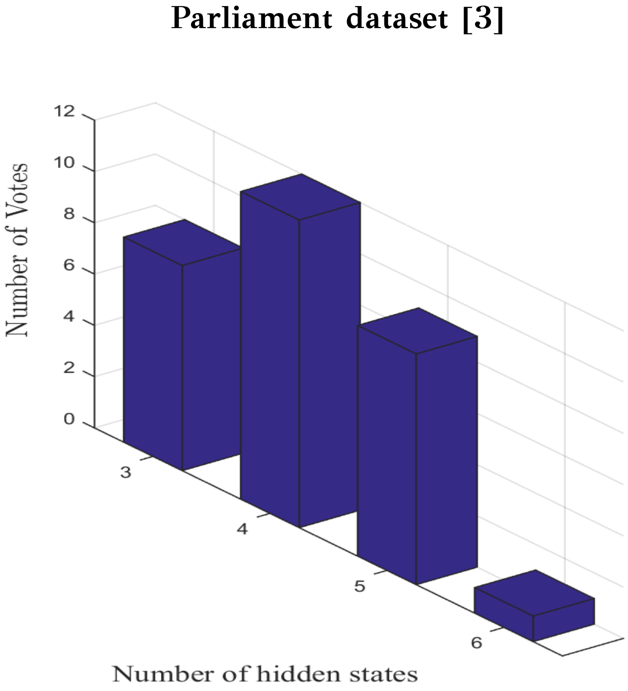


Figure 5.17: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the Parliament dataset [3] using synchronized and fusion features cues. The number of hidden states that do not appear in the horizontal axis received zero votes.

The confusion matrices of all methods for the dataset are given in Fig. 5.18. Comparing confusion matrices of the dataset's representation expressed by C3D and synchronized-fusion features we can note that each representation outlines and describes better different action classes. In particular, using the C3D representation all methods tend to identify better the neutral class whereas the use of synchronized-fusion representation helps the methods to recognize better the aggressive class. As a result, a fusion of those two representations may lead to higher accuracy score for the whole dataset.

The statistical information about the recognition accuracy for each action category and the overall score is summarized in Table 5.5. This table shows the mean accuracy accompanied by the standard deviation for each model. A comparison between Table 5.5 and Table 5.4 leads to the conclusion that C3D representation is better for this dataset attaining better accuracy results.

		Categories		
Method	**Overall**	**Aggressive**	**Friendly**	**Neutral**
SVM	75.45 ± 3.15	98.63 ± 2.9	65.56 ± 7.2	63.08 ± 6.4
CRF	82.12 ± 1.7	100 ± 0	75.65 ± 5.7	70.38 ± 10.5
HCRF	82.85 ± 2.5	94.52 ± 5.9	81.11 ± 6.3	72.31 ± 6.4
RI-HCRF	85.5 ± 3.1	95.89 ± 6.3	78.89 ± 4.6	83.08 ± 3.4

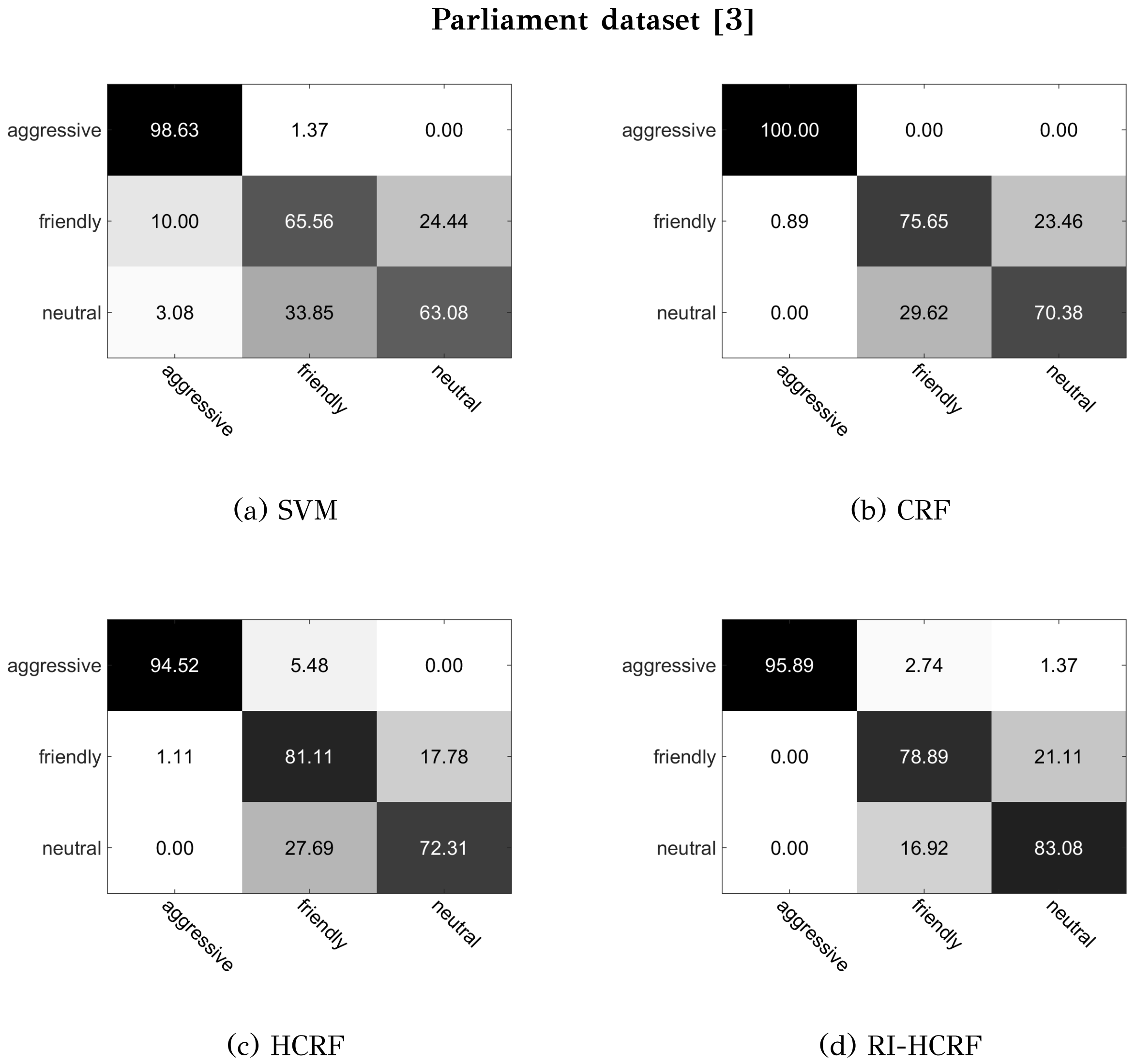
Table 5.5: Averaged recognition accuracies of all methods for the Parliament dataset [3] using synchronized and fusion features cues (mean ± st. dev.).

5.4.4 Evaluation on the TPI dataset

The TPI dataset will be described in two parts. In the first part, the results extracted from the C3D representation of the dataset will be presented while in the second one the results obtained using the synchronized-fusion feature representation will be discussed. In order to split the dataset into training and test set a 5-fold cross validation was used in both dataset's representations.

TPI dataset expressed by C3D features

The illustration of recognition accuracy for RI-HCRF and HCRF models is depicted in Fig. 5.19. We may observe that RI-HCRF model achieves a recognition rate of



(a) SVM

(b) CRF

(c) HCRF

(d) RI-HCRF

Figure 5.18: Confusion matrices for the classification results for the Parliament dataset [3] using synchronized and fusion features cues.

94.5% while the standard HCRF achieves a recognition rate of 91.53%. Also, both models in this configuration share the same variability pattern and for some hidden states the accuracy scores do not have a significant difference.

The results of the prediction of the optimal number of hidden states is shown in Fig. 5.20. The value of 9 turns out to be the most probable number of hidden state that should be used by the models. A prediction that is in fully agreement with the experimental results conducted in an exhaustive search approach. Also, we may observe that the number of candidates is lower compared to the previous results and this may be a result of the small number of features that is provided in this configuration of the dataset.

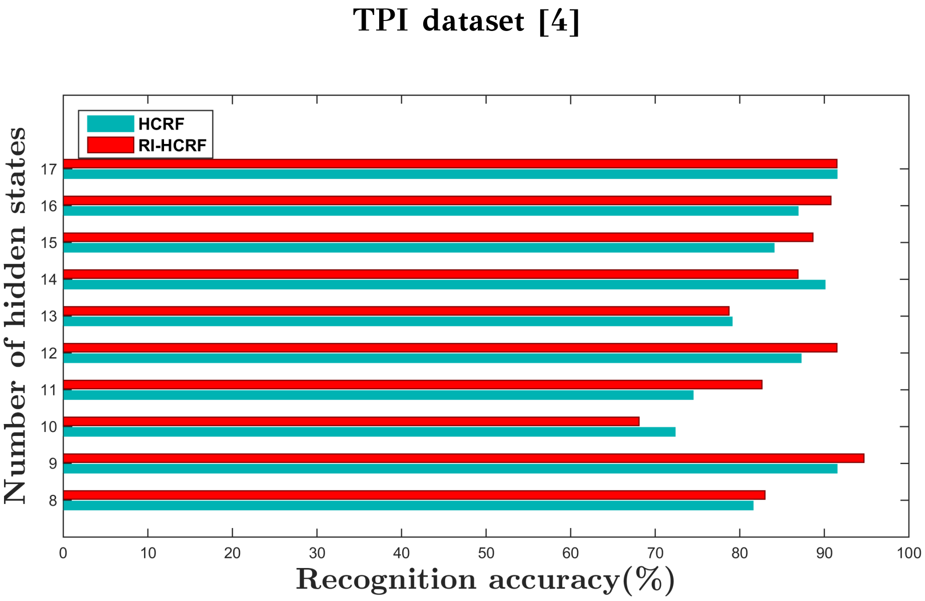


Figure 5.19: Classification accuracy with respect to the number of hidden states for the TPI dataset [4] using C3D features.

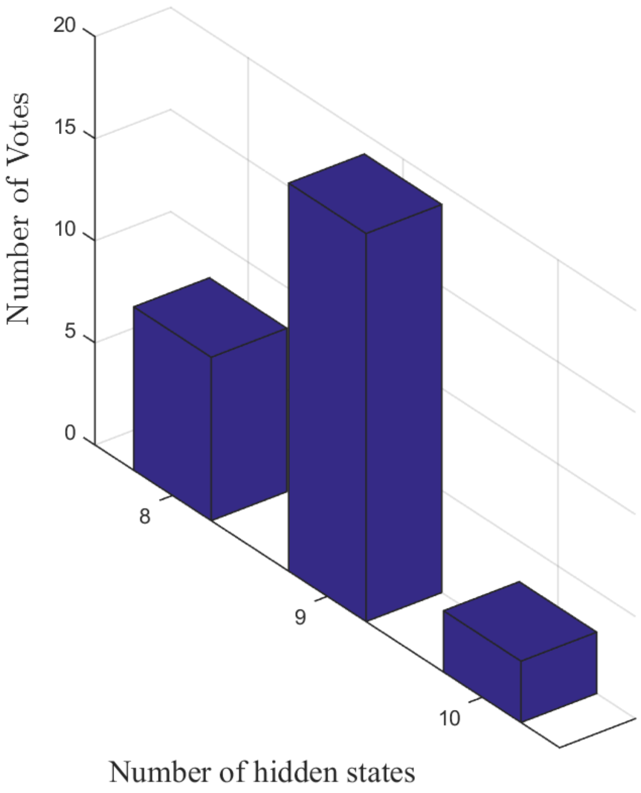
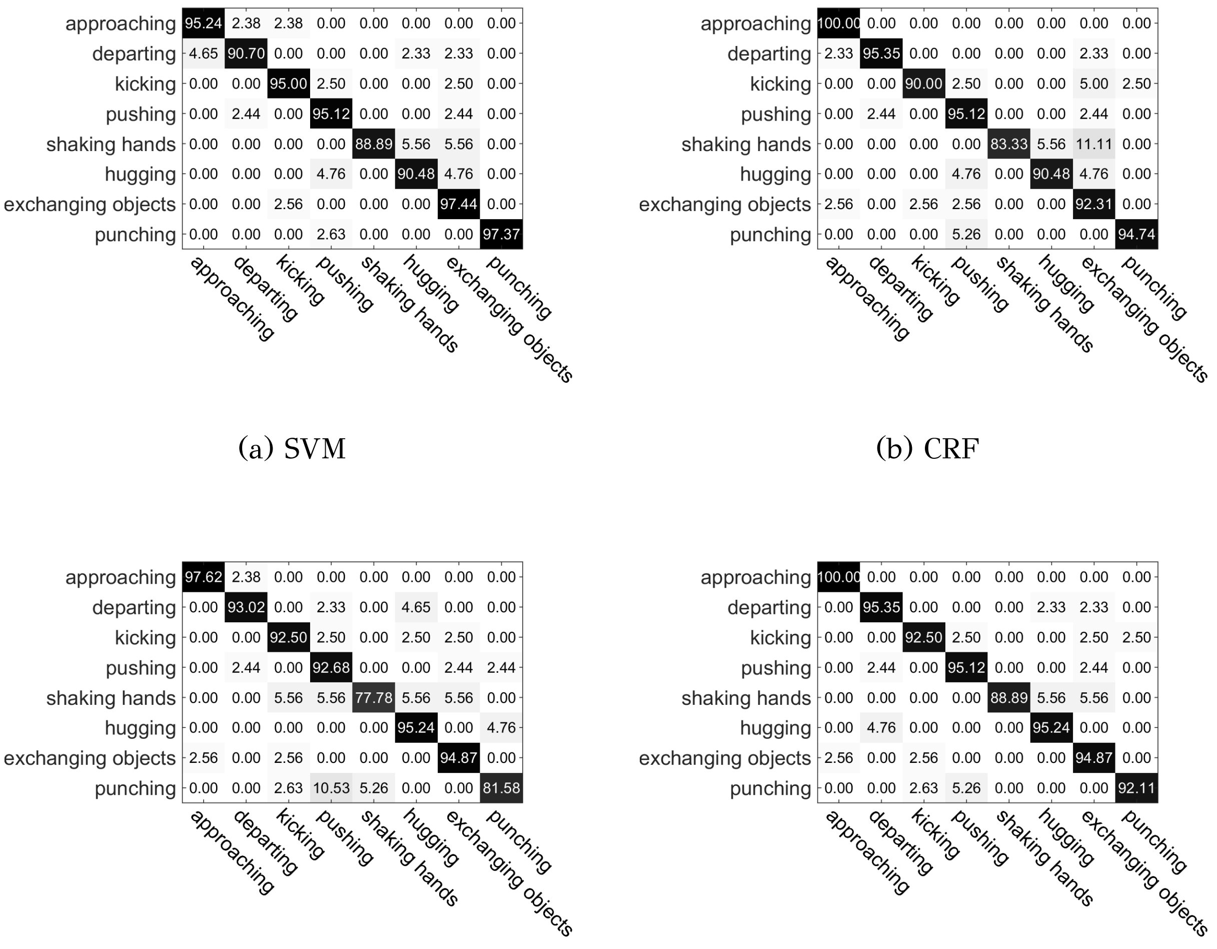


Figure 5.20: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the TPI dataset [4] using C3D features. The number of hidden states that do not appear in the horizontal axis received zero votes.

The confusion matrices of all methods for the dataset are given in Fig. 5.21. RI-HCRF and CRF classifies correctly reach 100% accuracy for the approaching action while the shaking hands action seems to be the less distinct action for all employed methods.

TPI dataset [4]



(a) SVM

(b) CRF

(c) HCRF

(d) RI-HCRF

Figure 5.21: Confusion matrices for the classification results for the TPI dataset [4] using C3D features.

The average recognition accuracy and its standard deviation for each model and each action category are presented in Table 5.6. It is also worth noting that RI-HCRF has the lowest deviation in the overall accuracy rate and in the majority of action categories. A comparison between the mean accuracy rate of the models indicates that RI-HCRF has the best results.

		Categories							
Method	**Overall**	**Approach**	**Depart**	**Kick**	**Push**	**Shake Hands**	**Hug**	**Exchange Objects**	**Punch**
SVM	94.27 ± 4.3	95.24 ± 9.9	90.70 ± 16.2	95.00 ± 11.2	95.12 ± 11.2	88.89 ± 16.2	90.48 ± 13.6	97.44 ± 6.4	97.37 ± 6.4
CRF	93.67 ± 3.5	100 ± 0	95.35 ± 9.9	90.0 ± 16.3	95.12 ± 9.9	83.33 ± 17.0	90.48 ± 13.6	92.31 ± 11.1	94.74 ± 7.3
HCRF	91.53 ± 3.7	97.62 ± 5.5	93.02 ± 6.3	92.5 ± 16.7	92.68 ± 10.1	77.78 ± 21.7	95.24 ± 11.2	94.87 ± 11.2	81.58 ± 11.8
RI-HCRF	94.75 ± 3.2	100 ± 0	95.35 ± 6	92.5 ± 11.1	95.12 ± 9.9	88.89 ± 16.2	95.24 ± 11.1	94.87 ± 11.1	92.11 ± 11.4

Table 5.6: Averaged recognition accuracies of all methods for the TPI dataset [4] using C3D features (mean ± st. dev.).

TPI dataset expressed by synchronized and fusion features

In Fig. 5.22 the bar representation of recognition accuracy with respect to the number of hidden variables is shown. RI-HCRF achieves a recognition accuracy rate of 89.37% while the standard HCRF achieves a recognition rate of 86.97%. The accuracy score for both models is decreased in this configuration of the dataset comparing to the previous one with C3D features. Furthermore, it seems that when the number of hidden states increases for both feature representations of this dataset the accuracy is getting high scores maybe due to the number actions and their intra-class variability.

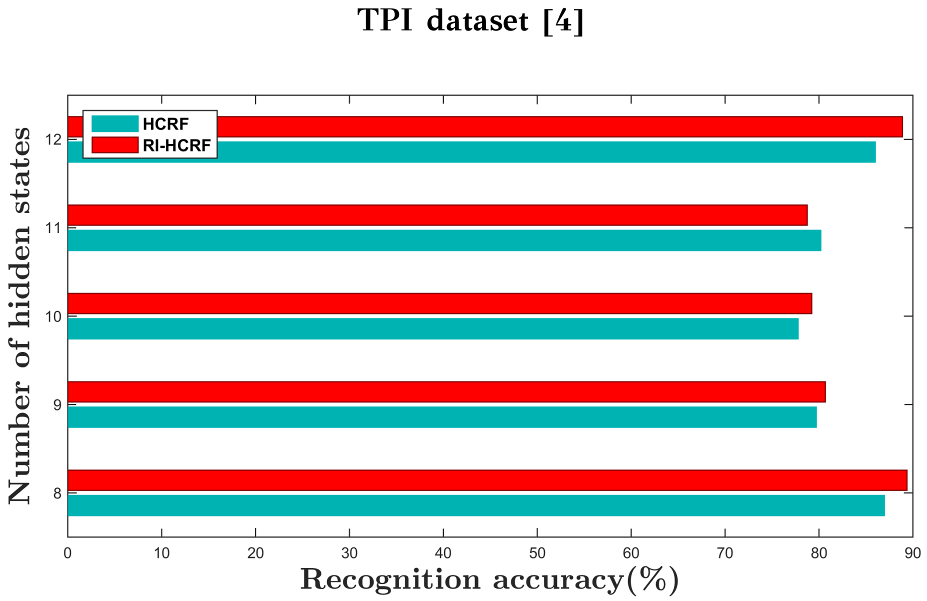


Figure 5.22: Classification accuracy with respect to the number of hidden states for the TPI dataset [4] using synchronized and fusion features cues.

The results of our method for the proposed optimal number of hidden states is given in Fig. 5.23 and indicates a model with 8 hidden variables. The highest scores for both HCRF and RI-HCRF are achieved using 8 hidden states and there is a total agreement between the optimal number of hidden variables derived from the prediction procedure and the exhaustive search depicted in (Fig. 5.22).

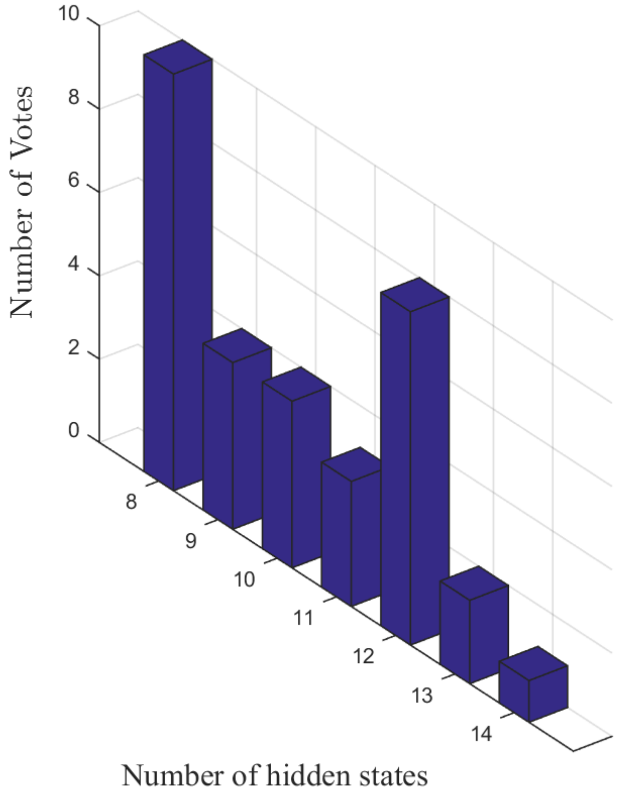
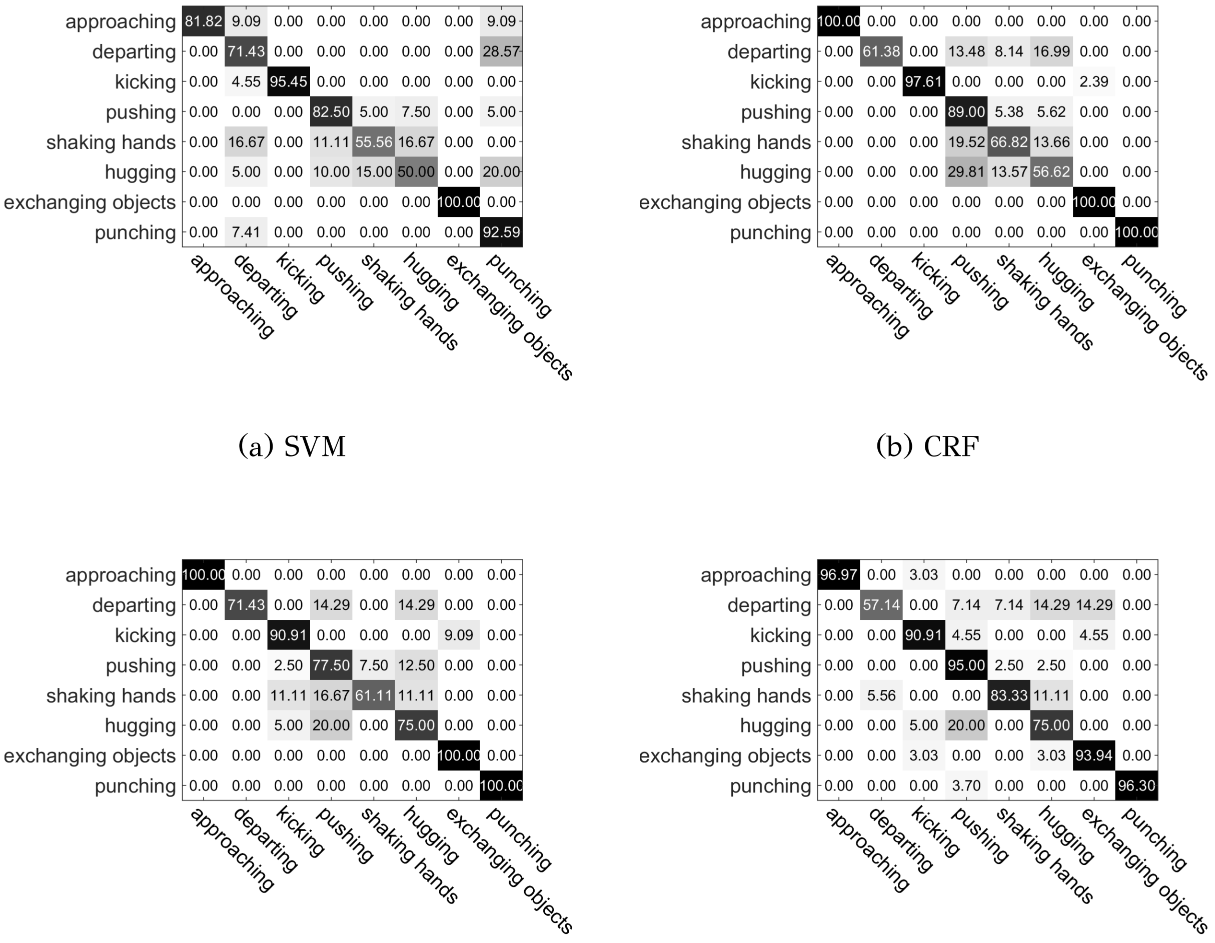


Figure 5.23: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the TPI dataset [4] using synchronized and fusion features cues. The number of hidden states that do not appear in the horizontal axis received zero votes.

The confusion matrices of all methods for the dataset are given in Fig. 5.24. The two of the four models (CRF and HCRF) are able to recognize perfectly not only the approaching action but also the exchange objects and punch actions for the TPI dataset expressed by synchronized and fusion features. On the other hand, RI-HCRF has satisfactory results and seems to recognize better actions like shaking hands and pushing that confuse all other models used in this work.

TPI dataset [4]



(a) SVM

	approaching	departing	kicking	pushing	shaking hands	hugging	exchanging objects	punching
approaching	81.82	9.09	0.00	0.00	0.00	0.00	0.00	9.09
departing	0.00	71.43	0.00	0.00	0.00	0.00	0.00	28.57
kicking	0.00	4.55	95.45	0.00	0.00	0.00	0.00	0.00
pushing	0.00	0.00	0.00	82.50	5.00	7.50	0.00	5.00
shaking hands	0.00	16.67	0.00	11.11	55.56	16.67	0.00	0.00
hugging	0.00	5.00	0.00	10.00	15.00	50.00	0.00	20.00
exchanging objects	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
punching	0.00	7.41	0.00	0.00	0.00	0.00	0.00	92.59

(b) CRF

	approaching	departing	kicking	pushing	shaking hands	hugging	exchanging objects	punching
approaching	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
departing	0.00	61.38	0.00	13.48	8.14	16.99	0.00	0.00
kicking	0.00	0.00	97.61	0.00	0.00	0.00	2.39	0.00
pushing	0.00	0.00	0.00	89.00	5.38	5.62	0.00	0.00
shaking hands	0.00	0.00	0.00	19.52	66.82	13.66	0.00	0.00
hugging	0.00	0.00	0.00	29.81	13.57	56.62	0.00	0.00
exchanging objects	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
punching	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(c) HCRF

	approaching	departing	kicking	pushing	shaking hands	hugging	exchanging objects	punching
approaching	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
departing	0.00	71.43	0.00	14.29	0.00	14.29	0.00	0.00
kicking	0.00	0.00	90.91	0.00	0.00	0.00	9.09	0.00
pushing	0.00	0.00	2.50	77.50	7.50	12.50	0.00	0.00
shaking hands	0.00	0.00	11.11	16.67	61.11	11.11	0.00	0.00
hugging	0.00	0.00	5.00	20.00	0.00	75.00	0.00	0.00
exchanging objects	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
punching	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(d) RI-HCRF

	approaching	departing	kicking	pushing	shaking hands	hugging	exchanging objects	punching
approaching	96.97	0.00	3.03	0.00	0.00	0.00	0.00	0.00
departing	0.00	57.14	0.00	7.14	7.14	14.29	14.29	0.00
kicking	0.00	0.00	90.91	4.55	0.00	0.00	4.55	0.00
pushing	0.00	0.00	0.00	95.00	2.50	2.50	0.00	0.00
shaking hands	0.00	5.56	0.00	0.00	83.33	11.11	0.00	0.00
hugging	0.00	0.00	5.00	20.00	0.00	75.00	0.00	0.00
exchanging objects	0.00	0.00	3.03	0.00	0.00	3.03	93.94	0.00
punching	0.00	0.00	0.00	3.70	0.00	0.00	0.00	96.30

Figure 5.24: Confusion matrices for the classification results for the TPI dataset [4] using synchronized and fusion features cues.

The summary of the above results is given by Table 5.7, where the average recognition accuracy and its standard deviation for each model and each action category are presented. A comparison between the mean accuracy rate of the models indicates that RI-HCRF has the best results.

		Categories							
Method	**Overall**	**Approach**	**Depart**	**Kick**	**Push**	**Shake Hands**	**Hug**	**Exchange Objects**	**Punch**
SVM	81.53 ± 6.8	81.82 ± 18.63	71.43 ± 47.1	95.45 ± 8.9	82.5 ± 11.1	55.56 ± 20.9	50.00 ± 17.6	100 ± 0	92.59 ± 10.6
CRF	87.29 ± 4.4	100 ± 0	61.38 ± 35.1	97.61 ± 5.4	89.00 ± 14.0	66.82 ± 30.1	56.62 ± 14.2	100 ± 0	100 ± 0
HCRF	86.97 ± 2.6	100 ± 0	71.43 ± 18.2	90.91 ± 12.5	77.5 ± 10.4	61.11 ± 25.2	75.0 ± 17.6	100 ± 0	100 ± 0
RI-HCRF	89.37 ± 2.7	96.97 ± 6.3	57.14 ± 38.1	90.91 ± 22.36	95 ± 6.8	83.33 ± 22.36	75 ± 17.68	93.94 ± 7.8	96.3 ± 7.4

Table 5.7: Averaged recognition accuracies of all methods for the TPI dataset [4] using synchronized and fusion features cues (mean ± st. dev.).

5.4.5 Evaluation on the THVI dataset

In same way like Parliament and TPI datasets, the TVHI dateset will be described in two parts:

1. TVHI dataset expressed by C3D features
2. TVHI dataset expressed by synchronized and fusion features,

where each part will be an evaluation of a different representation of the dataset. The test and the training sets are obtained by a 5-fold cross validation that was used in order to split the dataset.

TVHI dataset expressed by C3D features

The bar representation shown in FIg. 5.25 refers to the accuracy rate of the standard HCRF and our proposed model RI-HCRF with respect to the number of hidden variables. As we can see, the recognition accuracy rate of HCRF and R-HCRF is 93.0% and 93.5% respectively for each model using 7 hidden states Also, they seem to have the lowest accuracy using 4 hidden variables in Ri-HCRF case and using 9 hidden variables in the HCRF case. Apart from the first state, our model seems to outperform the standard HCRF.

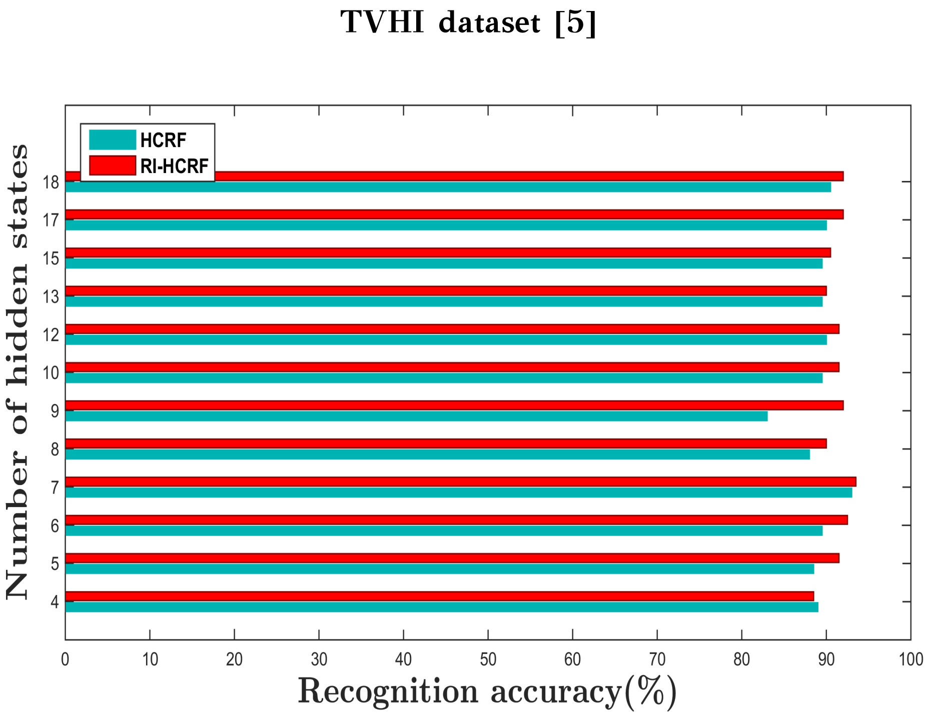


Figure 5.25: Classification accuracy with respect to the number of hidden states for the TVHI dataset [5] using C3D features.

The prediction results for the optimal number of hidden variables is given in Fig. 5.26, where the value of 7 is the most probable. The prediction and the exhaustive experimental results agree and point the same optimal number.

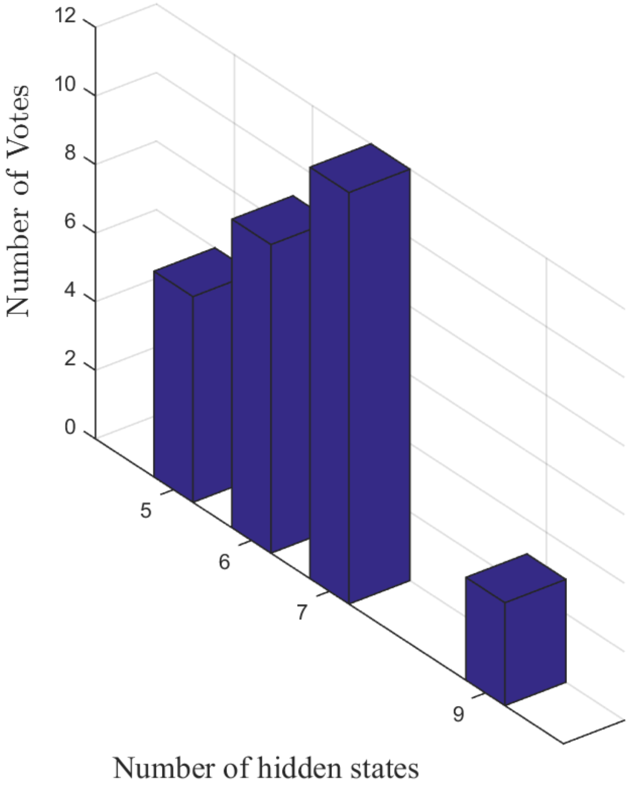


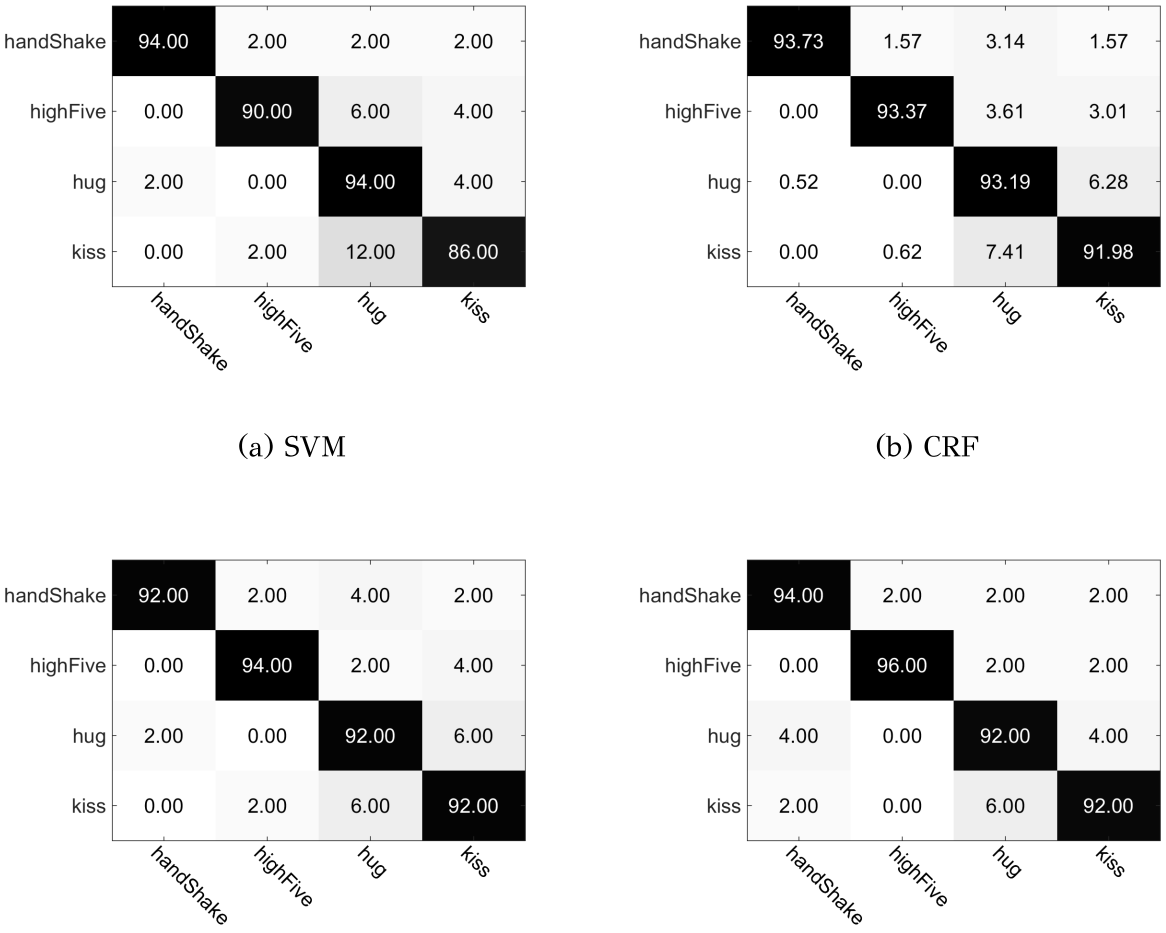
Figure 5.26: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the TVHI dataset [5] using C3D features. The number of hidden states that do not appear in the horizontal axis received zero votes.

The confusion matrix of each method used in this work for the TVHI dataset, is shown in Fig. 5.27. We can observe that the SVM is able to recognize better hand shake and hug actions, the CRF is able to recognize better the hand shake action while HCRF and RI-HCRF are able to recognize better the high five one.

Table 5.8 summarizes the statistical information about the evaluation of the TVHI dataset expressed by C3D features. This table presents for each model, the mean accuracy accompanied by the standard deviation of each action class and the overall score. A comparison between these scores shows that RI-HCRF achieves better results than all baseline methods.

TVHI dataset expressed by synchronized and fusion features

TVHI dataset [5]



	handShake	highFive	hug	kiss
handShake	94.00	2.00	2.00	2.00
highFive	0.00	90.00	6.00	4.00
hug	2.00	0.00	94.00	4.00
kiss	0.00	2.00	12.00	86.00

(a) SVM

	handShake	highFive	hug	kiss
handShake	93.73	1.57	3.14	1.57
highFive	0.00	93.37	3.61	3.01
hug	0.52	0.00	93.19	6.28
kiss	0.00	0.62	7.41	91.98

(b) CRF

	handShake	highFive	hug	kiss
handShake	92.00	2.00	4.00	2.00
highFive	0.00	94.00	2.00	4.00
hug	2.00	0.00	92.00	6.00
kiss	0.00	2.00	6.00	92.00

(c) HCRF

	handShake	highFive	hug	kiss
handShake	94.00	2.00	2.00	2.00
highFive	0.00	96.00	2.00	2.00
hug	4.00	0.00	92.00	4.00
kiss	2.00	0.00	6.00	92.00

(d) RI-HCRF

Figure 5.27: Confusion matrices for the classification results for the TVHI dataset [5] using C3D features.

		Categories			
Method	**Overall**	**Hand Shake**	**High Five**	**Hug**	**Kiss**
SVM	91.00 ± 5.5	94.00 ± 8.9	90.00 ± 7.0	94.00 ± 5.5	86.00 ± 15.1
CRF	92.82 ± 3.2	93.73 ± 7.8	93.37 ± 9.9	93.19 ± 5.4	91.98 ± 6.5
HCRF	93.0 ± 2.0	92.0 ± 5.4	94.0 ± 8.9	92.0 ± 4.4	92.0 ± 8.3
RI-HCRF	93.5 ± 2.2	94.0 ± 5.4	96.0 ± 5.4	92.0 ± 4.4	92.0 ± 8.3

Table 5.8: Averaged recognition accuracies of all methods for the TVHI dataset [5] using C3D features (mean ± st. dev.).

The illustration of the recognition accuracy of HCRF and RI-HCRF models with respect to the number of hidden variables is depicted in Fig. 5.28. By observing, we can notice that HCRF and RI-HCRF reach an accuracy rate of 99.5% and 100% using using 9 hidden states. Moreover, we can see that a large number of hidden variables seems to confuse the models than to enhance their accuracy results because when the number of hidden variables increases the accuracy of the models decreases rapidly.

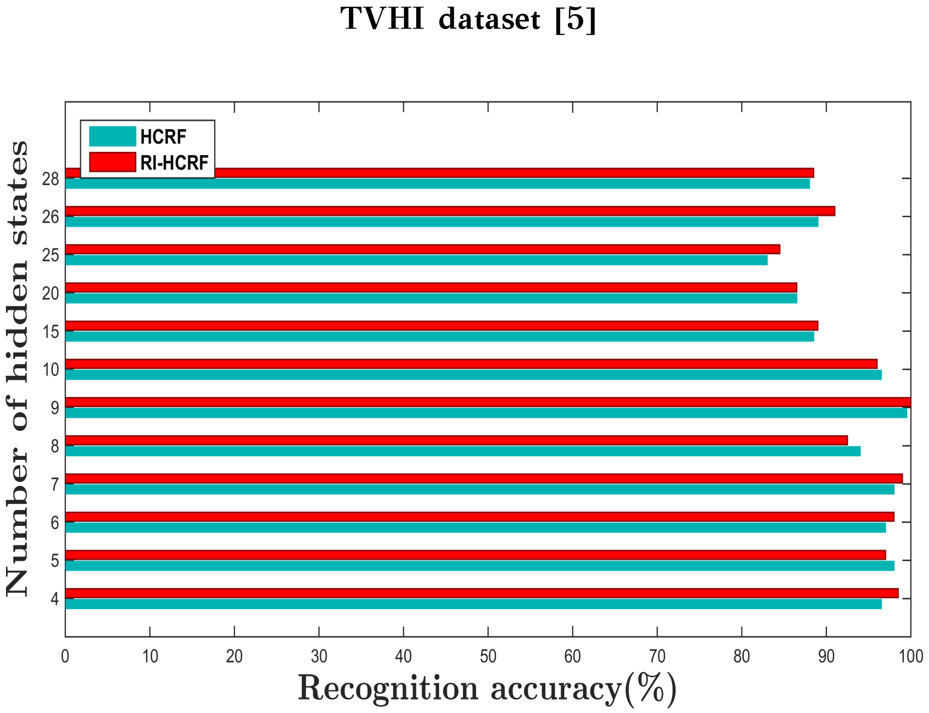


Figure 5.28: Classification accuracy with respect to the number of hidden states for the TVHI dataset [5] using synchronized and fusion features cues.

The optimal number of hidden variables according to the prediction results of our models is 9 (Fig. 5.29) and is identical to the optimal number of hidden variables obtained by the experimental results presented in Fig. 5.28.

The confusion matrices of all methods for the dataset are given in Fig. 5.30. The CRF and RI-HCRF models are able to recognize perfectly all the actions while HCRF and SVM are able to recognize perfectly only 3 and 2 actions respectively.

The average recognition accuracy accompanied by its standard deviation for each model and each action category are presented in Table 5.9. Comparing Tables 5.9 and 5.8 we can note that the synchronized and fusion features representation achieves better results in TVHI in contrast to Parliament and TPI for which the C3D representation has better accuracy scores.

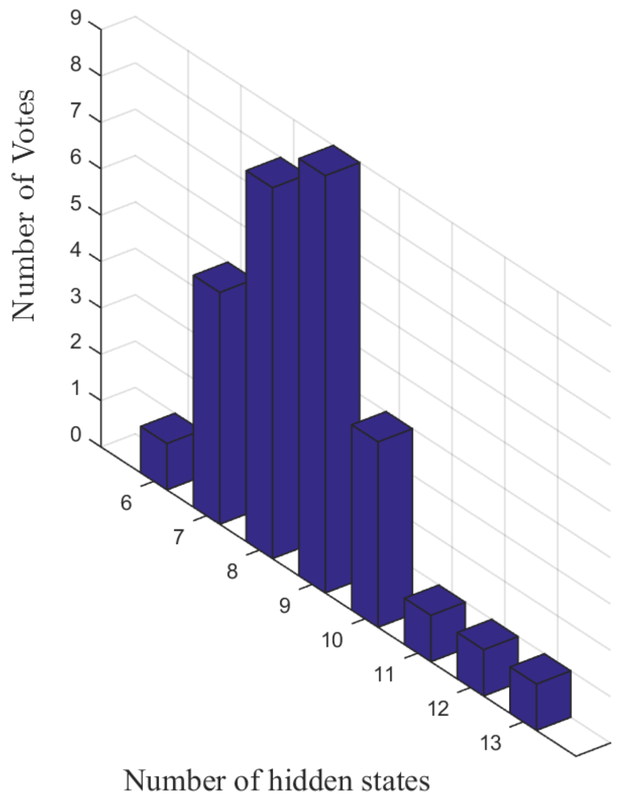
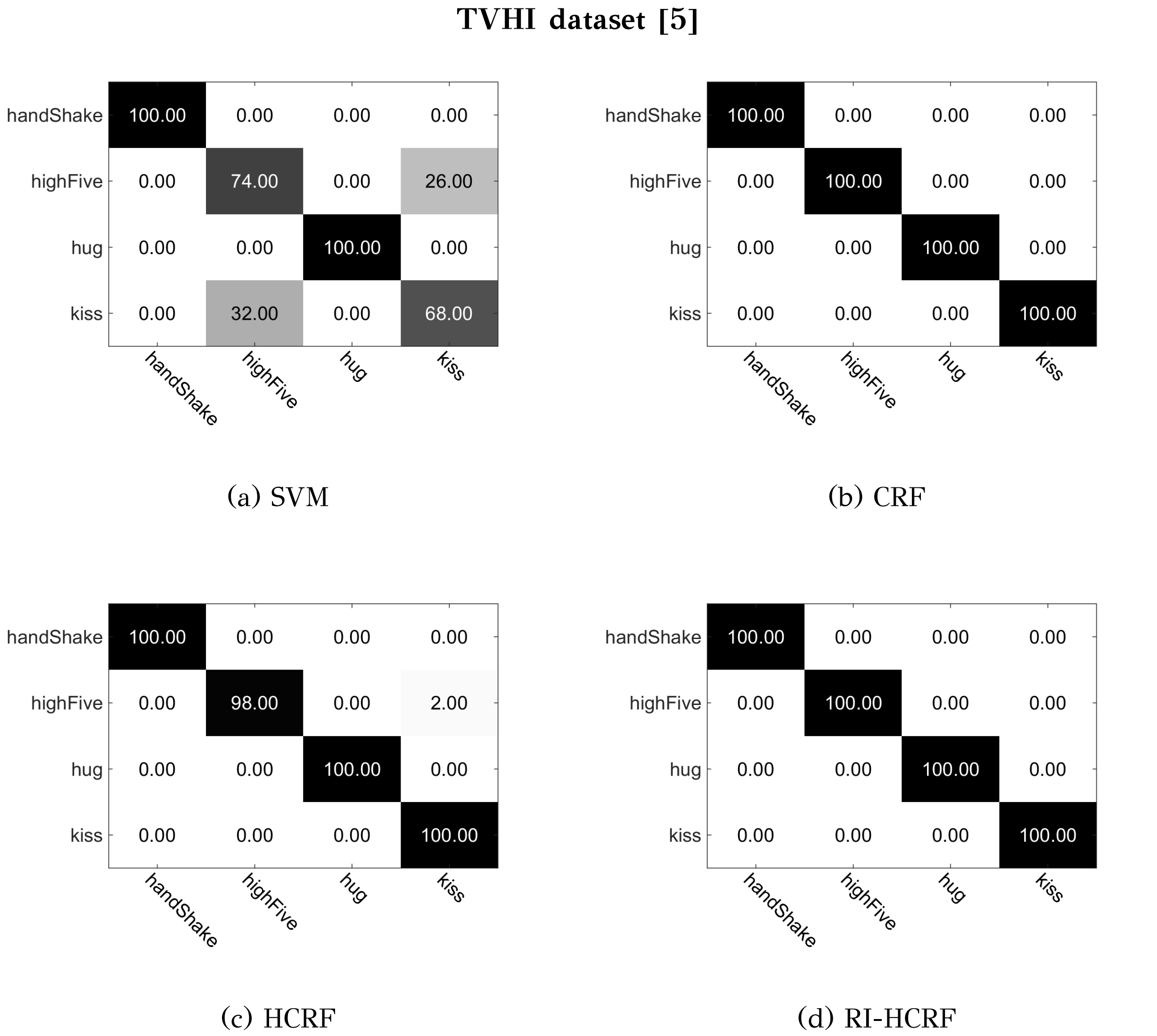


Figure 5.29: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the TVHI dataset [5] using synchronized and fusion features cues. The number of hidden states that do not appear in the horizontal axis received zero votes.

		Categories			
Method	**Overall**	**Hand Shake**	**High Five**	**Hug**	**Kiss**
SVM	85.50 ± 4.8	100 ± 0	74.00 ± 8.9	100.00 ± 0	68.00 ± 14.0
CRF	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
HCRF	99.5 ± 1.5	100 ± 0	98.0 ± 6.3	100 ± 0	100 ± 0
RI-HCRF	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0

Table 5.9: Averaged recognition accuracies of all methods for the TVHI dataset [5] using synchronized and fusion features cues (mean ± st. dev.).



TVHI dataset [5]

	handShake	highFive	hug	kiss
handShake	100.00	0.00	0.00	0.00
highFive	0.00	74.00	0.00	26.00
hug	0.00	0.00	100.00	0.00
kiss	0.00	32.00	0.00	68.00

(a) SVM

	handShake	highFive	hug	kiss
handShake	100.00	0.00	0.00	0.00
highFive	0.00	100.00	0.00	0.00
hug	0.00	0.00	100.00	0.00
kiss	0.00	0.00	0.00	100.00

(b) CRF

	handShake	highFive	hug	kiss
handShake	100.00	0.00	0.00	0.00
highFive	0.00	98.00	0.00	2.00
hug	0.00	0.00	100.00	0.00
kiss	0.00	0.00	0.00	100.00

(c) HCRF

	handShake	highFive	hug	kiss
handShake	100.00	0.00	0.00	0.00
highFive	0.00	100.00	0.00	0.00
hug	0.00	0.00	100.00	0.00
kiss	0.00	0.00	0.00	100.00

(d) RI-HCRF

Figure 5.30: Confusion matrices for the classification results for the TVHI dataset [5] using synchronized and fusion features cues.

5.4.6 Evaluation on the USAA dataset

In the USAA dataset we used a 5-fold cross validation to split the dataset into training and test sets and the C3D feature representation is used to express the dataset. The results of the recognition accuracy score with respect to the number of hidden variables is depicted in Fig. 5.31 for both HCRF and RI-HCRF models. The results show that RI-HCRF with a small difference, achieves a better accuracy score (91.81%) than HCRF (91.58%). Also, both models reach their best scores using 10 hidden states. By taking a closer look at the results we can observe that our proposed model keeps higher accuracy scores than the standard HCRF, in every number of hidden variables that was used.

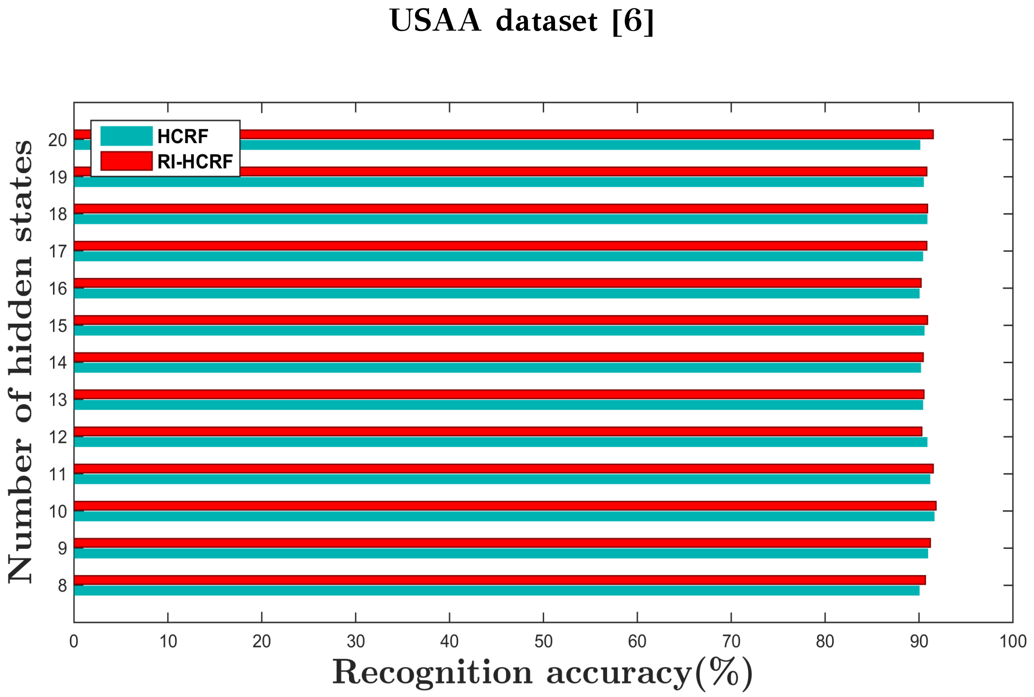


Figure 5.31: Classification accuracy with respect to the number of hidden states for the USAA dataset [6] using C3D features.

The prediction results for the optimal number of hidden variables is given in Fig. 5.32 and indicate the value of 10 to be the most probable number of hidden variables. The predicted number of hidden states is identical to the actual optimal number derived from the conducted experiments presented in Fig. 5.31.

The confusion matrix for each method used in this work for the USAA dataset, is shown in Fig. 5.33. We can observe that the social occasion that is recognized better from all the methods is birthday party while the wedding reception seems to have cause the most of confusion.

In Table 5.10 the statistical information about the evaluation of the USAA dataset expressed by C3D features is summarized. The mean accuracy accompanied by the standard deviation of each action class and the overall score are presented by the foregoing table. A comparison of the results leads to the conclusion that RI-HCRF outperforms all the baseline methods with the lowest standard deviation in the majority of the results.

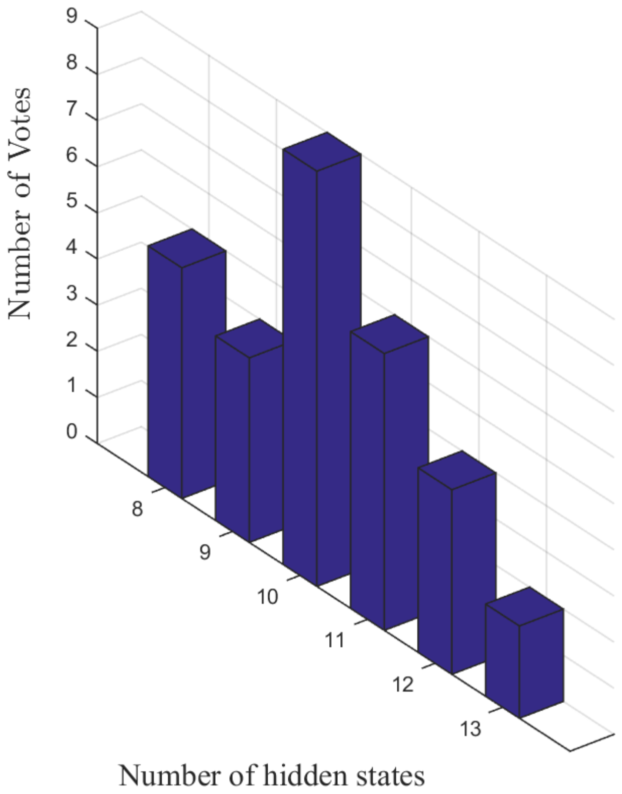
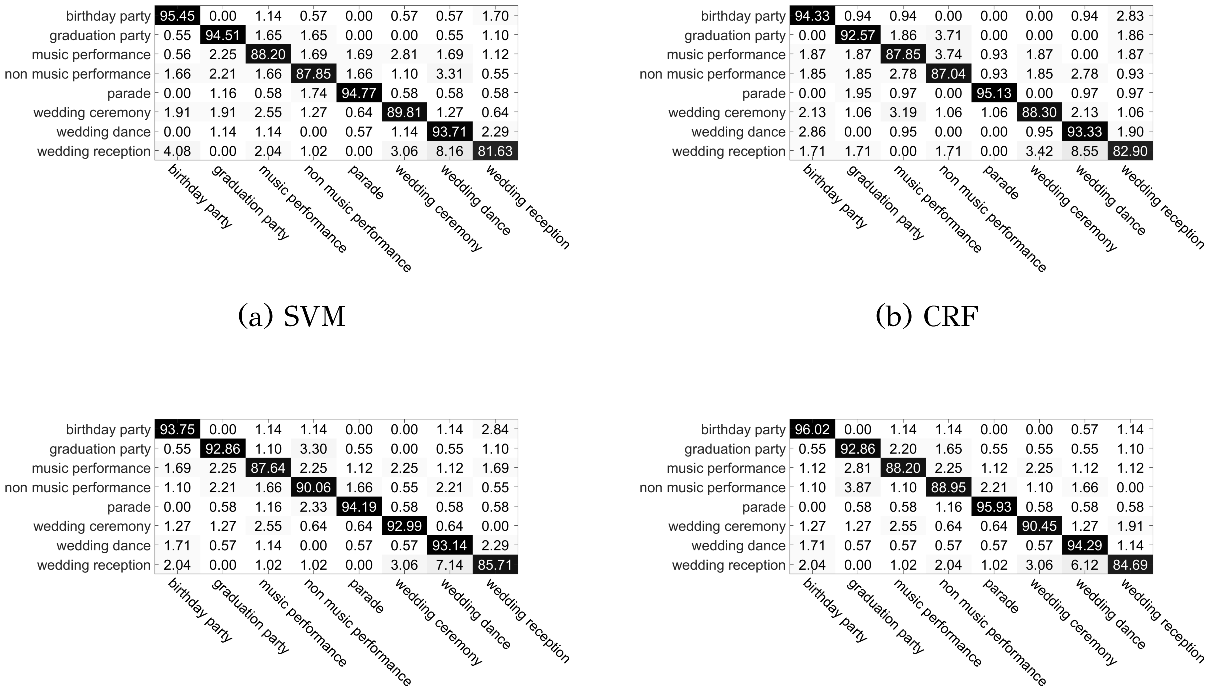


Figure 5.32: Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm for the USAA dataset [6] using C3D features. The number of hidden states that do not appear in the horizontal axis received zero votes.

		Categories							
Method	**Overall**	**Birthday**	**Graduation**	**Music**	**Non-music**	**Parade**	**Ceremony**	**Dance**	**Reception**
SVM	91.28 ± 1.4	95.45 ± 3.2	94.51 ± 5.2	88.20 ± 6.0	87.85 ± 5.8	94.77 ± 3.8	89.81 ± 8.7	93.71 ± 2.3	81.63 ± 9.0
CRF	90.62 ± 1.6	94.33 ± 2.7	92.57 ± 3.3	87.85 ± 6.9	87.04 ± 11.5	95.13 ± 3.4	88.30 ± 5.0	93.33 ± 4.3	82.90 ± 8.4
HCRF	91.58 ± 1.5	93.75 ± 5.3	92.86 ± 2.5	87.64 ± 5.8	90.06 ± 7.0	94.19 ± 3.6	92.99 ± 5.3	93.14 ± 4.3	85.71 ± 9.0
RI-HCRF	91.81 ± 1.1	96.02 ± 3.1	92.86 ± 2.5	88.20 ± 6.4	88.95 ± 9.8	95.93 ± 3.3	90.45 ± 3.3	94.29 ± 2	84.69 ± 8

Table 5.10: Averaged recognition accuracies of all methods for the USAA dataset [6] using C3D features (mean ± st. dev.).

USAA dataset [6]



	birthday party	graduation party	music performance	non music performance	parade	wedding ceremony	wedding dance	wedding reception
birthday party	95.45	0.00	1.14	0.57	0.00	0.57	0.57	1.70
graduation party	0.55	94.51	1.65	1.65	0.00	0.00	0.55	1.10
music performance	0.56	2.25	88.20	1.69	1.69	2.81	1.69	1.12
non music performance	1.66	2.21	1.66	87.85	1.66	1.10	3.31	0.55
parade	0.00	1.16	0.58	1.74	94.77	0.58	0.58	0.58
wedding ceremony	1.91	1.91	2.55	1.27	0.64	89.81	1.27	0.64
wedding dance	0.00	1.14	1.14	0.00	0.57	1.14	93.71	2.29
wedding reception	4.08	0.00	2.04	1.02	0.00	3.06	8.16	81.63

(a) SVM

	birthday party	graduation party	music performance	non music performance	parade	wedding ceremony	wedding dance	wedding reception
birthday party	94.33	0.94	0.94	0.00	0.00	0.00	0.94	2.83
graduation party	0.00	92.57	1.86	3.71	0.00	0.00	0.00	1.86
music performance	1.87	1.87	87.85	3.74	0.93	1.87	0.00	1.87
non music performance	1.85	1.85	2.78	87.04	0.93	1.85	2.78	0.93
parade	0.00	1.95	0.97	0.00	95.13	0.00	0.97	0.97
wedding ceremony	2.13	1.06	3.19	1.06	1.06	88.30	2.13	1.06
wedding dance	2.86	0.00	0.95	0.00	0.00	0.95	93.33	1.90
wedding reception	1.71	1.71	0.00	1.71	0.00	3.42	8.55	82.90

(b) CRF

	birthday party	graduation party	music performance	non music performance	parade	wedding ceremony	wedding dance	wedding reception
birthday party	93.75	0.00	1.14	1.14	0.00	0.00	1.14	2.84
graduation party	0.55	92.86	1.10	3.30	0.55	0.00	0.55	1.10
music performance	1.69	2.25	87.64	2.25	1.12	2.25	1.12	1.69
non music performance	1.10	2.21	1.66	90.06	1.66	0.55	2.21	0.55
parade	0.00	0.58	1.16	2.33	94.19	0.58	0.58	0.58
wedding ceremony	1.27	1.27	2.55	0.64	0.64	92.99	0.64	0.00
wedding dance	1.71	0.57	1.14	0.00	0.57	0.57	93.14	2.29
wedding reception	2.04	0.00	1.02	1.02	0.00	3.06	7.14	85.71

(c) HCRF

	birthday party	graduation party	music performance	non music performance	parade	wedding ceremony	wedding dance	wedding reception
birthday party	96.02	0.00	1.14	1.14	0.00	0.00	0.57	1.14
graduation party	0.55	92.86	2.20	1.65	0.55	0.55	0.55	1.10
music performance	1.12	2.81	88.20	2.25	1.12	2.25	1.12	1.12
non music performance	1.10	3.87	1.10	88.95	2.21	1.10	1.66	0.00
parade	0.00	0.58	0.58	1.16	95.93	0.58	0.58	0.58
wedding ceremony	1.27	1.27	2.55	0.64	0.64	90.45	1.27	1.91
wedding dance	1.71	0.57	0.57	0.57	0.57	0.57	94.29	1.14
wedding reception	2.04	0.00	1.02	2.04	1.02	3.06	6.12	84.69

(d) RI-HCRF

Figure 5.33: Confusion matrices for the classification results for the USAA dataset [6] using C3D features.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Video analysis has experienced a rapid rise of interest and research from the computer vision community due to the colossal number of videos provided, uploaded or stored. The main focus of this work is video content analysis involving graphical models. In this thesis we have introduced a novel approach that uses the multimedia content and its spatial and temporal dynamics for the recognition of human activities. This approach is an extension of the standard HCRF that aims to a better, robust and efficient recognition model.

Graphical models, which incorporate hidden variables in their structure, face the problem that the number of hidden variables need to be fixed by the user in advance. Our model, RI-HCRF, has been proposed in order to automatically infer the number of hidden variables from the input data. Also, a mixture model with three Student's t components is coupled to the RI-HCRF model as a prior to the parameters of the model because of its robustness to outlier values.

We have evaluated our model on six different benchmark datasets used for human action recognition. The challenge of these datasets is found in the resolution of videos, the heterogeneous background, the non static camera, the large number of clips and the high intra-class similarity between videos for certain classes. The evaluation of RI-HCRF has been done on several types of feature representations and has been compared to three others state of the art methods.

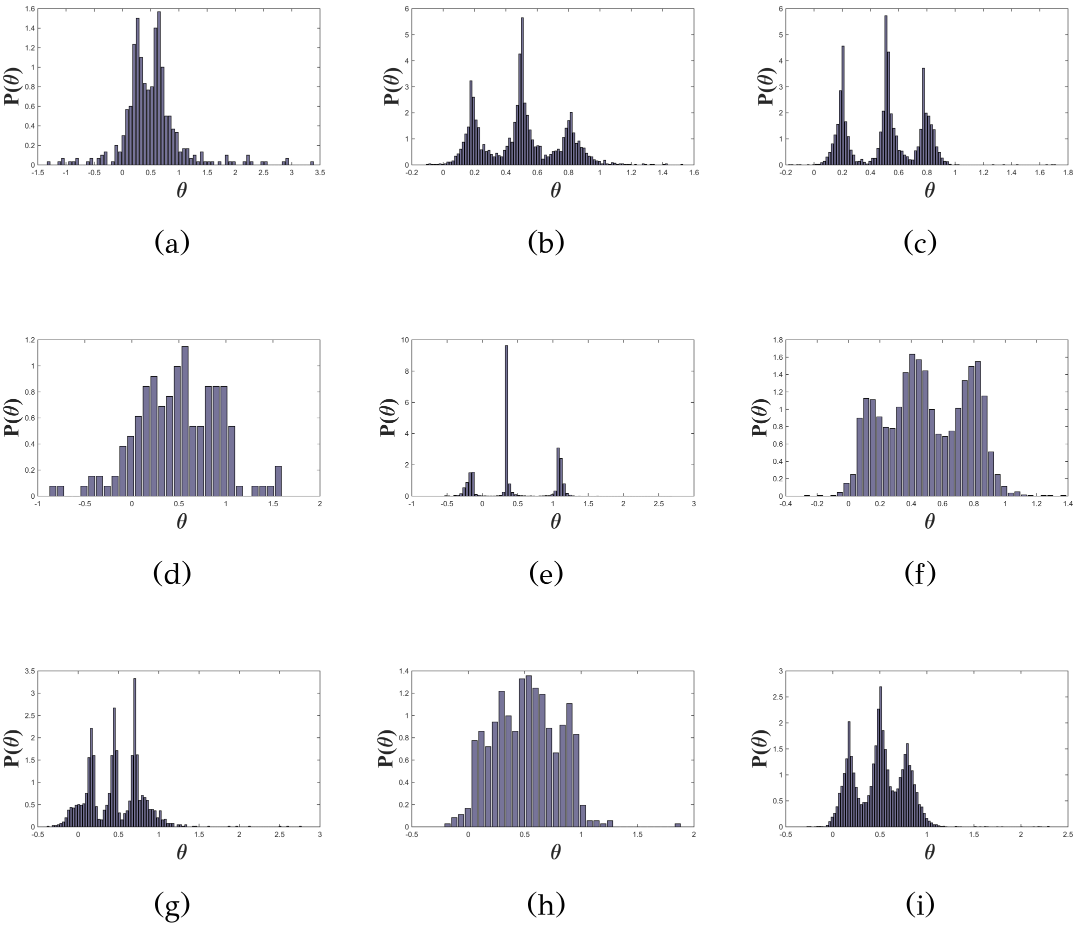
According to the experimental results, we conclude that the use of the incremental approach for the estimation of the hidden variables has been successful. Our method is able not only to estimate the optimal number of hidden variables correctly but also to mark a small neighborhood in which the optimal number is included.

In Fig. 6.1 the distribution of the trained parameters of a random fold for all datasets and their representations that were used in this work are shown. In the majority of the datasets the mixture distribution is visible and its components are distinct and not fully overlapping. Arm Gesture, Parliament and TVHI datasets, which use the fusion features representation are the datasets that the mixture model is not clear and maybe this is a result of the small number of parameters they use.

RI-HCRF has been proven a robust method with much potential and able to obtain satisfactory results, with better performance compared to the baseline models. However, in case of C3D representation of the datasets, its performance was comparable to the HCRF model. The reason of this result is the small set of features that C3D representation involves. Therefore, models discriminative power is similar to each other. Also, the C3D representation has been informative enough that a simpler method like a multi-class SVM was able to achieve good results.

Additionally, RI-HCRF and the other graphical models are suited to rich and overlapping features like synchronized and fusion ones. They seem to handle and understand better this kind of features in comparison with the standard multi-class SVM. Though, they achieve a lower overall classification accuracy than using 3D features derived from Convolutional Neural Networks.

The final conclusion from the conducted experiments is that each feature representation characterizes better different action category of the datasets due to the way and criterions the features were extracted. The number and the type of different modalities of the features (body joints location and orientation, intensity of voice and colour, space-time variations and texture) creates their heterogeneity and ability to describe each action from an alternative prospective.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 6.1: Distribution of the parameters θ of the RI-HCRF of a random fold for all datasets. (a) Arm Gesture dataset [1], (b) Weizmann dataset [2], (c) Parliament dataset [3] using C3D features, (d) Parliament dataset [3] using fusion features, (e) TPI dataset [4] using C3D features, (f) TPI dataset [4] using fusion features, (g) TVHI dataset [5] using C3D features, (h) TVHI dataset [5] using fusion features, (i) USAA dataset [6] using C3D features

6.2 Limitations and Future Work

One important limitation of our model is the dependence on the initialization of its parameters. The training of RI-HCRF is a challenging non-convex optimization problem and the possibility of getting stuck in local optima is high. The change of the objective function of the model from a non-convex to a convex one would be the best solution.

Also, the total performance of the model depends on the efficiency of the feature detector. The more efficient the detector is, the higher the recognition accuracy can be achieved. A small number of features, which that can hold enough information for the dataset can not only reduce the dimension of the video content but also the complexity of the model. It should be noted that the choice of features and the classification model should fit to each other. Thus, a fusion of C3D and STIP features may be a powerful combination for graphical models and especially for the structural modeling method RI-HCRF. A fusion of the dense STIP and the informative C3D features, could lead to a better representation of the datasets with higher recognition accuracy scores.

Despite the huge amount of research in the field of action recognition the recent decades, the need for a generic recognition system still exists. Building a generic model, able to classify correctly multiple action classes and successfully applicable to big data of the real wold systems, is a challenging task that should be further studied in the future.

Bibliography

[1] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1521–1527, 2006.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, pp. 1395–1402, 2005.

[3] M. Vrigkas, C. Nikou, and I. A. Kakadiadis, "Classifying behavioral attributes using conditional random fields," in *Proc. Hellenic Conference on Artificial Intelligence*, pp. 95–104, Springer, 2014.

[4] 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*, pp. 28–35, 2012.

[5] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, 2012.

[6] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *Proc. European Conference on Computer Vision*, pp. 530–543, Springer, 2012.

[7] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010.

[8] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Proc. European Conference on Computer Vision*, pp. 404–417, Springer, 2006.

[9] H. Lu, G. Fang, X. Shao, and X. Li, "Segmenting human from photo images based on a coarse-to-fine scheme," *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 889–899, 2012.

[10] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. European Conference on Computer Vision*, pp. 140–153, Springer, 2010.

[11] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005.

[13] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Spatiotemporal saliency for event detection and representation in the 3d wavelet domain: potential in human action recognition," in *Proc. ACM International Conference on Image and Video Retrieval*, pp. 294–301, 2007.

[14] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. ACM International Conference on Multimedia*, pp. 357–360, 2007.

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, 2007.

[18] J. Lafferty, A. McCallum, F. Pereira, *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. International Conference on Machine Learning*, vol. 1, pp. 282–289, 2001.

[19] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[20] A. Quattoni, M. Collins, and T. Darrell, “Conditional random fields for object recognition,” *Advances in Neural Information Processing Systems 17*, pp. 1097–1104, 2005.

[21] D. M. Gavrila, “The visual analysis of human movement: A survey,” *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[22] J. K. Aggarwal and Q. Cai, “Human motion analysis: A review,” in *Proc. Nonrigid and Articulated Motion Workshop*, pp. 90–102, IEEE, 1997.

[23] L. Wang, W. Hu, and T. Tan, “Recent developments in human motion analysis,” *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, 2003.

[24] 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, 2006.

[25] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, “A review of human activity recognition methods,” *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[26] D. Batra, T. Chen, and R. Sukthankar, “Space-time shapelets for action recognition,” in *Proc. IEEE Workshop on Motion and Video Computing*, pp. 1–6, 2008.

[27] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[28] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action mach a spatio-temporal maximum average correlation height filter for action recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[29] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” in *Proc. IEEE International Conference on Pattern Recognition*, vol. 3, pp. 32–36, 2004.

[30] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *Computer Vision and Image Understanding*, vol. 108, no. 3, pp. 207–229, 2007.

[31] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[32] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 3, pp. 710–719, 2005.

[33] P. Smith, N. da Vitoria Lobo, and M. Shah, "Temporalboost for event recognition," in *Proc. IEEE International Conference on Computer Vision*, vol. 1, pp. 733–740, IEEE, 2005.

[34] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," in *Proc. IEEE International Conference on Computer Vision*, pp. 1–8, IEEE, 2007.

[35] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[36] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. ACM International Conference on Multimedia*, pp. 357–360, 2007.

[37] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 379–385, 1992.

[38] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in *Proc. ACM International Conference on Ubiquitous Computing*, pp. 1–9, 2008.

[39] D. Hao Hu, S. J. Pan, V. W. Zheng, N. N. Liu, and Q. Yang, "Real world activity recognition with multiple goals," in *Proc. ACM International Conference on Ubiquitous Computing*, pp. 30–39, 2008.

[40] L.-P. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[41] T.-C. Liu, K.-C. Wang, A. Tsai, and C.-C. Wang, "Hand posture recognition using hidden conditional random fields," in *Proc. IEEE International Conference on Advanced Intelligent Mechatronics*, pp. 1828–1833, 2009.

[42] Y. Song, D. Demirdjian, and R. Davis, "Multi-signal gesture recognition using temporal smoothing hidden conditional random fields," in *Proc. IEEE International Conference and Workshops on Automatic Face & Gesture Recognition*, pp. 388–393, 2011.

[43] Y. Wang and G. Mori, "Learning a discriminative hidden part model for human action recognition," in *Proc. Annual Conference on Advances in Neural Information Processing Systems*, pp. 1721–1728, 2009.

[44] Y. Wang and G. Mori, "Max-margin hidden conditional random fields for human action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 872–879, 2009.

[45] J. Zhang and S. Gong, "Action categorization with modified hidden conditional random field," *Pattern Recognition*, vol. 43, no. 1, pp. 197–203, 2010.

[46] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1226–1233, 2012.

[47] Y. Song, L.-P. Morency, and R. Davis, "Multi-view latent variable discriminative models for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2120–2127, 2012.

[48] B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Affect analysis in natural human interaction using joint hidden conditional random fields," in *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2013.

[49] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Active privileged learning of human activities from weakly labeled samples," in *Proc. IEEE International Conference on Image Processing*, pp. 3036–3040, 2016.

[50] M. Vrigkas, C. Nikou, and I. Kakadiaris, “Identifying human behaviors using synchronized audio-visual cues,” 2016.

[51] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt, “Hidden conditional random fields for phone classification,” in *Proc. Interspeech*, pp. 1117–1120, 2005.

[52] Y.-H. Sung, C. Boulis, C. Manning, and D. Jurafsky, “Regularization, adaptation, and non-independent features improve hidden conditional random fields for phone classification,” in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 347–352, 2007.

[53] Y.-H. Sung, C. Boulis, and D. Jurafsky, “Maximum conditional likelihood linear regression and maximum a posteriori for hidden conditional random fields speaker adaptation,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4293–4296, 2008.

[54] Y.-H. Sung and D. Jurafsky, “Hidden conditional random fields for phone recognition,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 107–112, 2009.

[55] D. Yu, L. Deng, and A. Acero, “Hidden conditional random field with distribution constraints for phone classification,” in *Proc. Interspeech*, pp. 676–679, 2009.

[56] A. Vinel, T. M. T. Do, and T. Artieres, “Joint optimization of hidden conditional random fields and non linear feature extraction,” in *Proc. IEEE International Conference on Document Analysis and Recognition*, pp. 513–517, 2011.

[57] Y. Soullard and T. Artières, “Hybrid hmm and hcrf model for sequence classification,” in *Proc. European Symposium on Artificial Neural Networks*, 2011.

[58] S. Reiter, B. Schuller, and G. Rigoll, “Hidden conditional random fields for meeting segmentation,” in *Proc. IEEE International Conference on Multimedia and Expo*, pp. 639–642, 2007.

[59] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, “Efficient structured prediction with latent variables for general graphical models,” *International Conference on Machine Learning*, 2012.

[60] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction*, pp. 396–406, Springer, 2011.

[61] K. Bousmalis, L.-P. Morency, and M. Pantic, "Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition," in *Proc. IEEE International Conference and Workshops on Automatic Face & Gesture Recognition*, pp. 746–752, 2011.

[62] K. Bousmalis, S. Zafeiriou, L.-P. Morency, and M. Pantic, "Infinite hidden conditional random fields for human behavior analysis," *Transactions on Neural Networks and Learning Systems*, vol. 24, no. 1, pp. 170–177, 2013.

[63] K. Bousmalis, S. Zafeiriou, L.-P. Morency, M. Pantic, and Z. Ghahramani, "Variational hidden conditional random fields with coupled Dirichlet process mixtures," in *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 531–547, Springer, 2013.

[64] J. F. D. Saa and M. Cetin, "Hidden conditional random fields for classification of imaginary motor tasks from eeg data," in *Proc. IEEE European Signal Processing Conference*, pp. 171–175, 2011.

[65] S. Huh, R. Bise, M. Chen, T. Kanade, *et al.*, "Automated mitosis detection of stem cell populations in phase-contrast microscopy images," *Transactions on Medical Imaging*, vol. 30, no. 3, pp. 586–596, 2011.

[66] H. Kjellström, J. Romero, D. Martínez, and D. Kragić, "Simultaneous visual recognition of manipulation actions and manipulated objects," in *Proc. European Conference on Computer Vision*, pp. 336–349, Springer, 2008.

[67] M. A. El-Yacoubi, H. He, F. Roualdes, M. Selmi, M. Hariz, and F. Gillet, "Vision-based recognition of activities by a humanoid robot," *International Journal of Advanced Robotic Systems*, vol. 12, 2015.

[68] D. Xu, X. Wu, Y.-L. Chen, and Y. Xu, "Online dynamic gesture recognition for human robot interaction," *Journal of Intelligent & Robotic Systems*, vol. 77, no. 3-4, pp. 583–596, 2015.

[69] S. Oishi, Y. Kohari, and J. Miura, "Toward a robotic attendant adaptively behaving according to human state," in *Proc. IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1038–1043, 2016.

[70] G. Bouchard, "Bias-variance tradeoff in hybrid generative-discriminative models," in *Proc. IEEE International Conference on Machine Learning and Applications*, pp. 124–129, 2007.

[71] 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*, pp. 27–30, 2012.

[72] R. H. Byrd, J. Nocedal, and R. B. Schnabel, "Representations of quasi-Newton matrices and their use in limited memory methods," *Mathematical Programming*, vol. 63, no. 1, pp. 129–156, 1994.

[73] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[74] H. Permuter, J. Francos, and I. Jermyn, "A study of Gaussian mixture models of color and texture features for image classification and segmentation," *Pattern Recognition*, vol. 39, no. 4, pp. 695–706, 2006.

[75] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. IEEE International Conference on Pattern Recognition*, vol. 2, pp. 28–31, 2004.

[76] D. Reynolds, "Gaussian mixture models," *Encyclopedia of biometrics*, pp. 827–832, 2015.

[77] G. McLachlan and D. Peel, "Robust mixture modelling using the t distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 335–344, 2000.

[78] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[79] D. Demirdjian and T. Darrell, "3-d articulated pose tracking for untethered diectic reference," in *Proc. IEEE International Conference on Multimodal Interfaces*, p. 267, 2002.

[80] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.

[81] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

[82] V. N. Vladimir and V. Vapnik, "The nature of statistical learning theory," 1995.