Exploiting Privileged Information for Facial Expression Recognition

Michalis Vrigkas[1], Christophoros Nikou[1,2], Ioannis A. Kakadiaris[2]
[1]Department of Computer Science & Engineering, University of Ioannina, Ioannina 45110, Greece
mvrigkas@cs.uoi.gr, cnikou@cs.uoi.gr
[2]Computational Biomedicine Lab, University of Houston, 4800 Calhoun Rd., Houston, TX 77204, USA
ioannisk@uh.edu

Abstract

Most of the facial expression recognition methods consider that both training and testing data are equally distributed. As facial image sequences may contain information for heterogeneous sources, facial data may be asymmetrically distributed between training and testing, as it may be difficult to maintain the same quality and quantity of information. In this work, we present a novel classification method based on the learning using privileged information (LUPI) paradigm to address the problem of facial expression recognition. We introduce a probabilistic classification approach based on conditional random fields (CRFs) to indirectly propagate knowledge from privileged to regular feature space. Each feature space owns specific parameter settings, which are combined together through a Gaussian prior, to train the proposed t-CRF+ model and allow the different tasks to share parameters and improve classification performance. The proposed method is validated on two challenging and publicly available benchmarks on facial expression recognition and improved the state-of-the-art methods in the LUPI framework.

1. Introduction

Facial expression recognition has recently attracted much attention due to its applicability in several fields of biometrics, computer vision, and machine learning [30, 34]. Its applications may vary from video surveillance, driver and/or patient monitoring to human-machine interactions. Many facial expression recognition systems provide information about the personality and psychological state of a person. In real world, humans express their emotions as a combination of verbal and non-verbal multimodal cues such as gestures, facial expressions and auditory cues. Combining different modalities poses a great challenge on recognizing facial expressions [26, 27].

The multimodal nature of the problem requires the development of new learning techniques. Several approaches such as multi-task learning [19] and domain adaptation [10] have been proposed for dealing with multimodal problems. These approaches assume that the classifier is trained and tested on similar sets of data. However, exploiting the same type of information during training and testing may not always be possible due to data acquisition constraints. To this end, learning using privileged information (LUPI) [28] has been explored to cope with the inhomogeneity in training and testing information. The idea of privileged information is that one may have access to additional information about the training samples, which is not available during testing.

The LUPI framework emulates the human's perception of learning as it resembles the way that an educator teaches his/her students by providing additional knowledge, comments, explanations, or rewards in class, while the students latter are forced to solve problems without having access to this additional knowledge. In this context, the LUPI framework has also been used in several machine learning applications such as boosting [3], clustering [8], facial expression recognition [31] and textual description [25].

Learning using privileged information is a challenging task, since privileged information is only available during training and thus, an effective way of combining regular and privileged data is mandatory for recognizing the actual class label. In this context, the privileged information should be efficiently embedded into the classifier and estimate the model parameters. However, defining which information may be considered as privileged and which as regular is not an easy task as the problem is not straightforward [24], while the lack of informative data or the presence of misleading information may influence the performance of the model by introducing bias. There are several types of information that may be used as privileged. For example, in emotion recognition audio features may constitute a reasonable factor for understanding emotional states. Also, binary attributes such as facial action units may be used as auxiliary information for recognizing facial expressions.

In this work, we address these limitations by introducing a novel probabilistic model, which incorporates the LUPI

978-1-5090-1869-7/16/$31.00 ©2016 IEEE

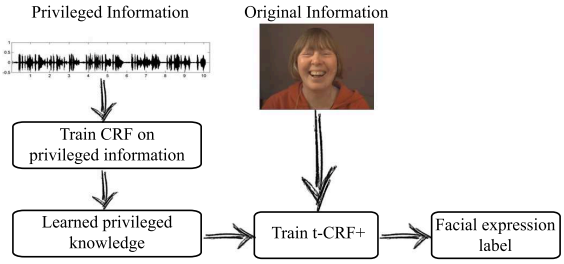

Figure 1. An overview of the proposed framework.

paradigm into a unified framework for recognizing facial expressions and affective states of a person. We propose an efficient method to indirectly transfer the knowledge from privileged to the original feature space using conditional random fields (CRFs) [13], called transfer-CRF+ (t-CRF+). Specifically, the privileged information is provided as additional input to our model through a two step classification process. We first train a standard CRF model on the privileged data and encode the ability of privileged information to distinguish between different class labels into the model weights. The learned privileged weights are then used to penalize the training process on the original feature space by learning the conditional probability distribution between the class labels and original observations. The penalty term encourages the model to assign larger weights to samples that have a good evidence to distinguish between classes both in privileged and original feature space and smaller weights to the contrary. In other words, the proposed model is able to enhance the classification accuracy by learning a better estimate of model parameters in the original feature space by transferring the knowledge from the privileged data. Figure 1 illustrates an overview of the proposed methodology.

The main contributions of our work can be summarized in the following points: (i) a new probabilistic classification scheme based on CRFs is proposed to improve the recognition of facial expressions and affective states of a person by gaining additional knowledge about the training data using privileged information; (ii) information transferring is used to keep only the relevant information between privileged and original feature space. Note that the proposed method is general and is not limited to the use of any specific form of privileged information, but rather it is general for any form of additional data.

The remainder of the paper is organized as follows: in Section 2, a review of the related work is presented. Section 3 presents the proposed t-CRF+ approach. In Section 4, experimental results are reported. Finally, conclusions are drawn in Section 5.

2. Related Work

Facial expression recognition methods may be divided into two broad categories, namely: frame-based and sequence-based methods [34]. Rather than recognizing static facial expressions, video-based methods are more natural related to human perception of understanding as facial events dynamically evolve over time. Thus, subtle facial expressions such as sadness or anger, that may efficiently be recognized from video sequences, may not be identifiable in static frames [1]. Another difference is that sequence-based approaches usually have smaller training and testing sets than frame-based methods.

In this context, Walecki *et al.* [29] proposed a variation of hidden conditional random fields (HCRF) [21] to model hidden dynamics of sequential facial expressions and automatically select the optimal model that can better discriminate between different facial expressions. The work of Dapogny *et al.* [6] was focused on bridging the gap between sequential and static classification of facial expressions by combining transition classifiers from both geometric and appearance features and fusing static and dynamic information from different time intervals. Lörincz *et al.* [14] used dynamic programming kernels with facial feature points for emotional expression recognition. All these methods assume pre-segmented facial expression sequences. As an alternative, Wu *et al.* [32] combined multiple instance learning and hidden Markov models (HMMs) to identify facial emotions from multiple peaks of expression.

Much research has also been focused in combining appearance features and the facial action coding system (FACS) [15, 20, 26]. Song *et al.* [26] studied the problem of facial expression recognition in partially labeled data under the terms of sparsity and compressed sensing. However, manually annotating facial expressions with facial action units (AU) is very time consuming due to the large amount of data. To this end, Girard *et al.* [9] varied the number of training data in facial expression recognition systems to estimate the optimal amount of input data required to automatically detect AUs and improve classification accuracy.

Multi-modal approaches based not only on facial but also on audio features have recently gained much popularity for recognizing affective facial states of a person [5, 23]. Usually, facial expressions are accompanied with vocal expressions that enhance the feeling of a person about the corresponding situation such as pain or surprise. Relying on that fact, Meng and Bianchi-Berthouze [17] proposed a hybrid method based on HMMs to classify audio/visual affective expressions through a multistage classification approach. Ramirez *et al.* [22] proposed a modification of HCRFs, called latent-dynamic conditional random fields (LDCRFs) to model the interaction between different high-level modalities (audio and video) using late fusion to classify facial expressions on affect. Although their method performs well, the lack of an intrinsic audio-visual relationship estimation limits the recognition problem. Song *et al.* [27] exploited the sparsity of temporal motion patterns to iden-

tify audio/visual facial emotions through sparse codebook learning.

The LUPI paradigm was first introduced by Vapnik and Vashist [28] as a new classification setting to model a real world learning process (i.e., teacher-student learning relationship) in a max-margin framework, called SVM+. However, SVM+ is computationally more expensive than standard SVM since it requires a contemporary estimation of the loss function for original and privileged space. Wang and Ji [31] exploited privileged information to recognize facial expressions by proposing two different loss functions, which can be adapted to any classifier. The first model encoded privileged information as an additional feature during training, while the second approach considered that privileged information can be represented as secondary labels. Serra-Toro *et al.* [24] proved that successfully selecting information that can be treated as privileged is not a straightforward problem. The choice of different types of privileged information in the context of an object classification task implemented in a max-margin scheme was also discussed by Sharmanska *et al.* [25]. Both original and privileged features were considered of equivalent difficulty for recognizing the true class. Finally, Yang and Patras [33] trained conditional random regression forests for detecting facial features, where the privileged information was used for choosing proper split functions at some randomly selected internal node.

3. Leaning with privileged information

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

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

3.1. t-CRF+ model formulation

Our method uses CRFs, which are defined by a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (see Fig. 2), as the probabilistic framework for modeling the facial expressions of a subject in a single image or video. During training, a classifier and the mapping from observations to the label set for the different configurations are learned. In testing, a probe sequence is classified into its respective state using belief propagation (BP) [12].

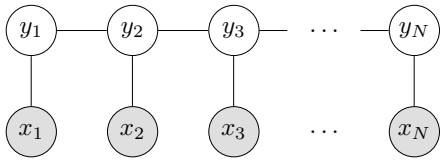

Figure 2. Graphical representation of the chain structure CRF model. The grey nodes are the observed features (x_i) and the white nodes are unknown labels (y_i), respectively.

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

$$p(y|\mathbf{x};\mathbf{w}) = \exp\left(E(y|\mathbf{x};\mathbf{w}) - A(\mathbf{w})\right) , \tag{1}$$

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

$$A(\mathbf{w}) = log \sum_{y'} \exp\left(E(y'|\mathbf{x};\mathbf{w})\right) . \tag{2}$$

Different sufficient statistics $E(y|\mathbf{x}, \mathbf{x}^*;\mathbf{w})$ in (1) define different distributions. In the general case, sufficient statistics consist of indicator functions for each possible configuration of unary and pairwise terms:

$$E(y|\mathbf{x};\mathbf{w}) = \sum_{j\in\mathcal{V}} \Phi(y_j, \mathbf{x}_j;\boldsymbol{\theta}) + \sum_{j,k\in\mathcal{E}} \Psi(y_j, y_k;\boldsymbol{\omega}) , \tag{3}$$

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

The unary potential is expressed by:

$$\Phi(y_j, \mathbf{x}_j;\boldsymbol{\theta}) = \sum_j \sum_{a\in\mathcal{Y}} \boldsymbol{\theta}^\top \mathbb{1}(y_j = a)\mathbf{x}_j , \tag{4}$$

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

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

$$\Psi(y_j, y_k;\boldsymbol{\omega}) = \sum_{a,b\in\mathcal{Y}} \sum_{\ell} \boldsymbol{\omega}_\ell \mathbb{1}(y_j = a)\mathbb{1}(y_k = b) . \tag{5}$$

Index ℓ corresponds to the number of the pairwise potentials.Note that the CRF model keeps a transition matrix for each label.

3.2. Parameter learning and inference

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

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w}) - \frac{\|\mathbf{w}\|^2}{2\sigma^2} \,. \tag{6}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}; \mathbf{w})$ and quantifies how well the distribution in Eq. (1) defined by the parameter vector $\mathbf{w}$ matches the labels y. The second term is a Gaussian prior with variance σ^2 and works as a regularizer.

Our work is based on the intuition that privileged information is more informative than the ordinary information and thus, learning on privileged data may improve the classification. The proposed t-CRF+ model relies on the idea that instead of jointly learning the ordinary and privileged information, we first train an ordinary CRF on the privileged feature space $\mathcal{X}^*$, and then we exploit the obtained knowledge to improve the performance on the target feature space $\mathcal{X}$, for which training data are always available during training and testing.

To achieve the knowledge transfer, we penalize the loss function of the standard CRF model with an additional term that corresponds to a Gaussian prior with zero mean and variance σ_p^2. Thus, the loss function in Eq. (6) is modified to encode the knowledge transfer from privileged to original feature space:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w_o}) - \frac{\|\mathbf{w_p} - \mathbf{w_o}\|^2}{2\sigma_p^2} - \frac{\|\mathbf{w_o}\|^2}{2\sigma_o^2} \,, \tag{7}$$

where $\mathbf{w_o}$ and $\mathbf{w_p}$ are the model parameters when training in the original and the privileged feature space, respectively. In Eq. (7), the parameters $\mathbf{w_o}$ and $\mathbf{w_p}$ should be of equal length and this is achieved using canonical correlation analysis (CCA) [11] as a preprocessing step. The parameters σ_p^2 and σ_o^2 are tuning parameters that control the degree of influence of the privileged and the original information, respectively.

Figure 3 illustrates the graphical representation of the proposed t-CRF+ model. The t-CRF+ model is parameterized by two hyper-parameters $\mathbf{w_p}$ and $\mathbf{w_o}$. In this case, the privileged information is indirectly transferred for learning the baseline CRF model through the learned prediction function for each training instance in the privileged space. The privileged parameters $\mathbf{w_p}$ are used in the original conditional log-likelihood function to influence the values of the parameters in the original feature space.

Algorithm 1 Transferring knowledge from $\mathcal{X}^*$ to $\mathcal{X}$ using t-CRF+

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

1: Perform canonical correlation to make the dimensions of $\mathcal{X}$ and $\mathcal{X}^*$ equal.
2: Train a standard CRF on the privileged data $(\mathbf{x}^*, y)$ using Eq. (6) and estimate models' parameters $\mathbf{w_p}$.
3: Train a CRF on the original feature space $(\mathbf{x}, y)$ using Eq. (7) to transfer the knowledge from the privileged to the original feature space.
4: Obtain final labels using Eq. (8).

The degree of influence the privileged information may have upon the original information depends on the degree of evidence for each privileged weight. The smallest the values of the privileged weights $\mathbf{w_p}$ are, the smallest the influence of privilege data also is. The opposite occurs when samples with larger privileged weights $\mathbf{w_p}$ may contribute more heavily through the Gaussian prior in Eq. (7) and thus, the privileged knowledge may have greater effect on the finally parameter learning. This process can be viewed as selection process, where the most informative data in the privileged space contribute to the classification of the true label.

In our implementation, the loss function in Eq. (7) is optimized using a gradient-descent optimization method. More specifically, we used the limited-memory BFGS (LBFGS) method [18] to minimize the negative log-likelihood of the data.

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

$$y = \arg\max_{y \in \mathcal{Y}} p(y|\mathbf{x}; \mathbf{w}) \,. \tag{8}$$

The marginal probability is obtained by applying the BP algorithm [2] using the graphical model as depicted in Fig. 2. The main steps of the proposed t-CRF+ classification model are summarized in Algorithm 1.

4. Experimental results

To show the ability of the proposed t-CRF+ method to generalize, we compared it with several state-of-the-art methods for two different computer vision applications, namely emotional facial recognition, and facial expression recognition, with different type of privileged information for each problem. For the first problem, we used the AVEC 2011 dataset [23] and for the second we used the extended Cohn-Kanade (CK+) dataset [15].

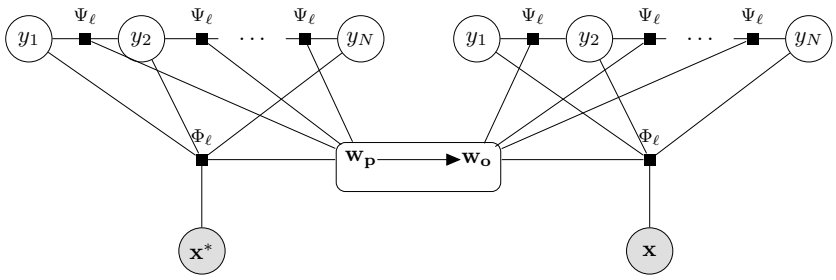

Figure 3. Proposed t-CRF+ model. First, a standard chain structure CRF model is trained on the privileged feature space ($\mathcal{X}^*$) with parameters $\mathbf{w_p}$. Then, the privileged knowledge is transferred to the original feature space ($\mathcal{X}$). The square nodes correspond to the unary and pairwise potentials, which are conditioned on their hyper-parameters $\mathbf{w_p}$ and $\mathbf{w_o}$, respectively.

4.1. Datasets

AVEC 2011 audio/visual challenge dataset [23]: This dataset consists of 95 sequences of upper body video segments at resolution of 780×580 at 49.979 fps while the audio was recorded at 48 kHz, and is part of the SEMAINE corpus [16]. The AVEC 2011 dataset consists of 31 videos for training, 32 videos for validation, and 32 videos for testing, annotated with four affective labels such as activation, expectation, power, and valence. As original features, we used the pre-computed video features provided by the dataset, and the privileged information was selected to be the provided audio features, which were obtained from various low-level descriptors. Due to the large amount of data and relatively high feature dimensionality for this dataset, we followed the same strategy as proposed by Schuller *et al.* [23] for sub-sampling the data and reducing the feature dimension.

Cohn-Kanade (CK+) dataset [15]: This dataset describes facial expressions such as anger, disgust, fear, happiness, sadness, surprise, and contempt. All facial expressions are expressed by the facial action coding system (FACS) [7], which describes all possible facial expressions as a combination of action units (AU), extracted from each participant, to identify their emotional state. It consists of 593 video sequences of 123 subjects captured from the neutral face to the peak expression. Since FACS are coded only at the peak frame, we only considered the peak frame in our experiments. For this dataset, the original features were selected to be the 68 tracked facial landmarks obtained by active appearance models [4] and the privileged information was selected to be the 17 annotated action units, all provided by the database creators.

4.2. Baseline approaches

We compared the proposed method with several baseline methods that may or may not use privileged information. First, we used SVM+ [28], which consists of optimizing the hyperplane parameters such that it can minimize the probability of incorrect classifications and increase the convergence rate. The second baseline is the rank transfer SVM+ (rt-SVM+) [25], which exploits a max-margin technique to transfer knowledge from the privileged to the original feature space. Finally we compared with the method of Wang and Ji [31], which exploits a loss inequality regularization (LIR) to address the sensitiveness of the loss function against the inequality constraints.

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

4.3. Model selection

The L_2 regularization scale terms σ_p and σ_o were set to 10^k, with $k \in \{-3, \ldots, 3\}$. The optimal parameters for all baseline methods were selected using cross validation, and the best parameters or parameter sets were used to retrain the model. Finally, our model in Eq. (7) was trained with a maximum of 400 iterations for the termination of the LBFGS minimization method.

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

4.4. Results and discussion

In the first set of experiments, we assessed the impact of privileged information to recognize affective states of emotional audio and video dyadic interactions between human participants using the AVEC 2011 dataset [23], and we also

Dataset	Regular	Privileged	Accuracy (%)	AUC (%)
AVEC 2011 [23]	visual	✗	60.5	85.7
	audio	✗	59.6	83.1
	visual+audio	✗	60.7	70.6
	visual	audio	**70.7**	**91.2**
CK+ [15]	facial lnd	✗	85.4	91.9
	AU	✗	85.1	92.5
	facial lnd + AU	✗	85.9	93.4
	facial lnd	AU	**93.6**	**99.3**

Table 1. Comparison of feature combinations for classifying facial expressions and affective states on AVEC 2011 [23], and CK+ [15] datasets. The crossmark indicates the absence of privileged information during training.

trained the proposed model to the CK+ dataset [15] for recognizing facial expressions. For the evaluation of the proposed method we used the classification accuracies and the area under the ROC curve (AUC), which compares the true positive against the false positive rate. The benefit of using robust privileged information along with conventional data instead of using each modality separately or both modalities as regular information is shown Table 1. For the classification, we used a standard CRF model and compared it with the proposed t-CRF+ method. We may observe that for both datasets, if only privileged information is used as regular features for classification both the classification accuracy and the AUC are lower than when using only the regular information for the classification task. However, these results are relatively similar to each other, which leads to the conclusion that finding proper privileged information is not always a straightforward procedure. Moreover, the proposed classification scheme performs better than all other approaches. These results demonstrate that the t-CRF+ model can successfully exploit the privileged information to improve the recognition accuracy.

In the second set of experiments, the proposed approach was compared with several state-of-the-art methods, that may or may not use privileged information for both datasets. The results are presented in Table 2. The results indicate that our approach improved the classification accuracy and the AUC. On AVEC 2011, we significantly managed to increase the classification accuracy by approximately 10% and the AUC by 20% with respect to CRF and SVM, which do not employ privileged information, as our approach achieves very high recognition accuracy for this dataset (70.7%). The improvement of our method compared to the methods that also employ privileged information is high. Furthermore, our method outperforms by approximately 7% in recognition accuracy and by 5% in AU the rt-SVM+, which also employs transferring of privileged information. Accordingly, for the CK+ dataset, the improvement against the state-of-the-art methods is also high and almost 8% higher accuracy with respect to the achieved by rt-SVM+ and 6% higher when compared to SVM+ and LIR

Method	AVEC 2011		CK+	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Methods without privileged information				
SVM [2]	57.3 ± 0.1	73.7 ± 0.3	84.8 ± 0.1	87.3 ± 0.1
CRF [13]	60.7 ± 0.8	70.6 ± 0.4	85.9 ± 0.6	93.4 ± 0.1
Methods with privileged information				
rt-SVM+ [25]	63.6 ± 0.1	86.3 ± 0.1	85.7 ± 0.1	88.4 ± 0.2
SVM+ [28]	59.6 ± 0.1	65.7 ± 0.1	87.7 ± 0.1	85.6 ± 0.1
LIR [31]	49.3 ± 0.1	67.2 ± 0.2	87.3 ± 0.8	85.5 ± 0.1
t-CRF+	$\mathbf{70.7} \pm 0.3$	$\mathbf{92.9} \pm 0.1$	$\mathbf{93.6} \pm 0.7$	$\mathbf{99.3} \pm 0.0$

Table 2. Comparison of the classification accuracies and the area under the ROC curve ($\%$) for the AVEC 2011 [23] and the CK+ [15] datasets (mean $\pm$ standard deviation).

Method	AVEC 2011		CK+	
	Accuracy	**AUC**	**Accuracy**	**AUC**
SVM [2]	0.0174	0.0383	0.0257	0.0089
CRF [13]	0.0435	0.0145	0.0390	0.0851
rt-SVM+ [25]	0.0269	0.7683	0.0361	0.0001
SVM+ [28]	0.0035	0.0062	0.0776	0.0026
LIR [31]	0.0043	0.0054	0.0666	0.0025

Table 3. p-values of the proposed method for the AVEC 2011 [23] and the CK+ [15] datasets.

methods. We may also observe that for this dataset, the AUC values achieved by the proposed t-CRF+ model are very high and close to the ideal classifier. In general, the significantly high increase in all evaluation indices by our model indicates the strength of the proposed method.

In order to provide a statistical evidence of the recognition results, we computed the p-values of the obtained results with respect to the compared methods. The null hypothesis was defined as: the mean performances (accuracies or AUC) of the proposed model are equal to the state-of-the-art methods; and the alternative hypothesis was defined as: the mean performances (accuracies or AUC) of the proposed model are higher than those of the state-of-the-art methods. For the assessment of the statistical significance, we used paired t-tests with statistical significance threshold $p < 0.05$ for all experiments. The resulted p-values for both datasets are reported in Table 3. According to these results, we conclude that for both datasets the the null hypothesis is rejected as the p-values were less than the significance level of 0.05, and thus, the improvements obtained by our model are statistically significant and not due to chance.

The corresponding ROC curves for both datasets are depicted in Fig. 4. The red dotted diagonal line corresponds to complete random guess. The intersection of the ROC curve for each method with the black diagonal line, corresponds to the equal error rate (EER). We may see that for the AVEC 2011 dataset the proposed method has the lowest EER (0.1141) and for the CK+ the EER is 0.0726, which is smaller than the state-of-the-art methods.

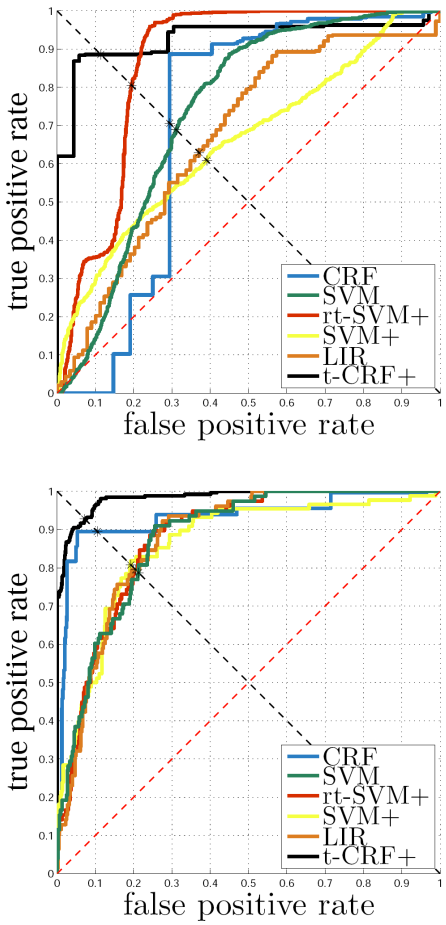

Figure 4. ROC curves for AVEC 2011 [23] (top row) and CK+ [15] (bottom row) datasets (best viewed in color).

Finally, the classification performance of the proposed method against the baseline methods for each class separately on both datasets is depicted in Fig. 5. We may observe that for AVEC 2011 in three out of four classes the proposed t-CRF+ method has the highest accuracy. However, for the valence class the standard CRF model performs slightly better, but still our method outperforms the rest of the state-of-the-art. For the CK+ dataset, the classification accuracy on four classes is perfect (100%), but for the classes sadness and surprise the proposed method performs worse than the baseline methods, mostly because some action units are hard to detect.

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

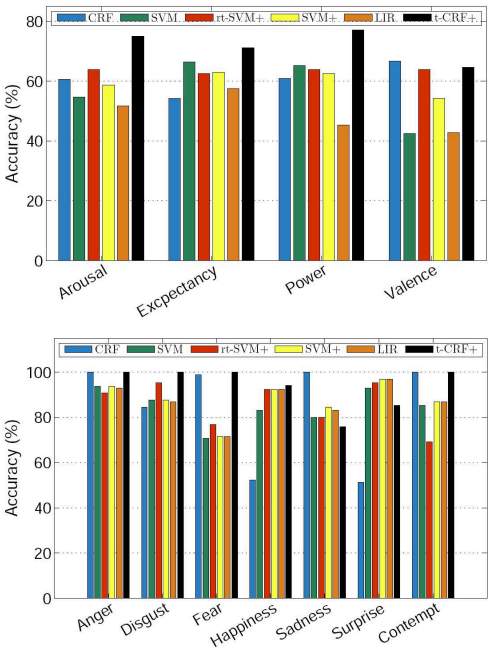

Figure 5. Comparison of recognition performance accuracies (%) of each class for AVEC 2011 [23] (top row) and CK+ [15] (bottom row) datasets (best viewed in color).

5. Conclusion

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

In the future, we plan to evaluate our method on multiple and heterogeneous sources of privileged information and assess the quality of the privileged information in other classification problems in biometrics.

Acknowledgments. This research was funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors. The work of C. Nikou was supported by the European Commission (H2020-MSCA-IF-2014), under grant agreement No 656094.

References

[1] Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 16(5):403–410, 2005.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] J. Chen, X. Liu, and S. Lyu. Boosting with side information. In *ACCV*, 2012.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[5] A. Cullen and N. Harte. Late integration of features for acoustic emotion recognition. In *EUSIPCO*, 2013.

[6] A. Dapogny, K. Bailly, and S. Dubuisson. Dynamic facial expression recognition by joint static and multi-time gap transition classification. In *FG*, 2015.

[7] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, 2002.

[8] J. Feyereisl and U. Aickelin. Privileged information for data clustering. *Information Sciences*, 194(0):4–23, 2012.

[9] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. D. la Torre. How much training data for facial action unit detection? In *FG*, 2015.

[10] R. Gopalan, L. Ruonan, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[11] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, 16(12), 2004.

[12] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Trans. on Image Processing*, 16(11):2649–2661, 2007.

[13] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICVML*, 2001.

[14] A. Lörincz, L. A. Jeni, Z. Szabó, J. F. Cohn, and T. Kanade. Emotional expression classification using time-series kernels. In *CVPRW*, 2013.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[16] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *ICME*, 2010.

[17] H. Meng and N. Bianchi-Berthouze. Affective state level recognition in naturalistic facial and vocal expressions. *IEEE Trans. on Cybernetics*, 44(3):315–328, 2014.

[18] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, 2nd edition, 2006.

[19] S. Parameswaran and K. Q. Weinberger. Large margin multitask metric learning. In *NIPS*, 2010.

[20] P. Perakis, T. Theoharis, and I. A. Kakadiaris. Feature fusion for facial landmark detection. *Pattern Recognition*, 47(9):2783–2793, 2014.

[21] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852, 2007.

[22] G. A. Ramirez, T. Baltrušaitis, and L. P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *ICACII*, 2011.

[23] B. Schuller, M. Valstar, F. Eybenn, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011-the first international audio/visual emotion challenge. In *ICACII*, 2011.

[24] C. Serra-Toro, V. J. Traver, and F. Pla. Exploring some practical issues of SVM+: Is really privileged information that helps? *Pattern Recognition Letters*, 42(0):40–46, 2014.

[25] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. In *ICCV*, 2013.

[26] Y. Song, D. McDuff, D. Vasisht, and A. Kapoor. Exploiting sparsity and co-occurrence structure for action unit recognition. In *FG*, 2015.

[27] Y. Song, L. P. Morency, and R. Davis. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In *ICMI*, 2013.

[28] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

[29] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *FG*, 2015.

[30] S. Wang and Q. Ji. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Trans. on Affective Computing*, 6(4):410–430, 2015.

[31] Z. Wang and Q. Ji. Classifier learning with hidden information. In *CVPR*, 2015.

[32] C. Wu, S. Wang, and Q. Ji. Multi-instance hidden Markov model for facial expression recognition. In *FG*, 2015.

[33] H. Yang and I. Patras. Privileged information-based conditional regression forest for facial feature detection. In *FG*, 2013.

[34] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.