Pattern Analysis and Appications manuscript No.
(will be inserted by the editor)

Human Activity Recognition Using Robust Adaptive Privileged Probabilistic Learning

Michalis Vrigkas · Evangelos Kazakos · Christophoros Nikou · Ioannis A. Kakadiaris

Received: date / Accepted: date

Abstract In this work, a supervised probabilistic approach is proposed that integrates the learning using privileged information (LUPI) paradigm into a hidden conditional random field (HCRF) model, called HCRF+, for human action recognition. The proposed model employs a self-training technique for automatic estimation of the regularization parameters of the objective function. Moreover, the method provides robustness to outliers by modeling the conditional distribution of the privileged information by a Student's t-density function, which is naturally integrated into the HCRF+ framework. The proposed method was evaluated using different forms of privileged information on four publicly available datasets. The experimental results demonstrate its effectiveness concerning the-state-of-the-art in the LUPI framework using both hand-crafted and deep learning-based features extracted from a convolutional neural network.

Keywords Hidden conditional random fields · learning using privileged information · human activity recognition · Student's t-distribution

This work has been co-funded by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EDK-04517) and by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund.

Michalis Vrigkas
Department of Communication and Digital Media, University of Western Macedonia, Kastoria, Greece

Christophoros Nikou
Department of Computer Science and Engineering, University of Ioannina, Ioannina, Greece

Evangelos Kazakos
Department of Computer Science, University of Bristol, Bristol, UK

Ioannis A. Kakadiaris
Department of Computer Science, University of Houston, Houston, TX, USA

1 Introduction

Recent advances in computer vision such as video surveillance, human-machine interactions, and semantic multimodal analysis [5, 53, 61, 22] rely on machine learning techniques trained on large-scale human-annotated datasets. However, training data may not always be available during testing, and learning using privileged information (LUPI) [34, 36] has been used to tackle this problem. The idea behind privileged information is that one may have access to additional information about the training samples, which is not available during testing.

Consequently, classification models may often suffer from "structure imbalance" between training and testing data, which may be represented by the LUPI paradigm. The LUPI technique simulates a real-life learning condition, when a student learns from his/her teacher, where the latter provides the student with additional knowledge, comments, explanations, or rewards in class. Subsequently, the student should be able to face any problem related to what he/she has learned without the help of the teacher.

The problem of human activity understanding using privileged knowledge is on its own a difficult task. Since privileged information is only available during training, one should combine both regular and privileged information into a unified classifier to predict the true class label. However, it is quite difficult to identify the most useful information to be used as privileged as the lack of informative data or the presence of misleading information may influence the performance of the model.

We address these issues by presenting a probabilistic approach, based on hidden conditional random fields (HCRFs) [45], called HCRF+. The proposed method can learn human activities by exploiting additional information about the input data, that may reflect on natural or auxiliary properties about classes and members of the classes of the training data (Fig. 1). This information is used for training purposes only but not for predicting the true classes (where, in general, this information is missing).

In particular, the proposed HCRF+ method differentiates from previous approaches [58], which may also use the LUPI paradigm, by incorporating privileged information in a supervised probabilistic manner, which facilitates the training process by learning the conditional probability distribution between human activities and observations. We also introduce a novel technique for automatic estimation of the optimal regularization parameters for the learning process. The method is adaptive as the regularization parameters are computed from the training data through a self-training procedure. It is worth noting that the proposed methodology is not limited to the use of a specific form of privileged information, but it is general and may handle any form of additional data.

Moreover, our method can efficiently manage dissimilarities in input data, which may correspond to noise, missing data, or outliers, using a Student's t-distribution to model the conditional probability of the privileged information. Such dissimilarities may harm the classification accuracy and lead to excessive sensitivity when input data is insufficient or contains large intra-class variations. In particular, the use of Student's t-distribution is justified by the property that it has heavier tails than a standard Gaussian distribution, thus providing robustness to outliers [43].

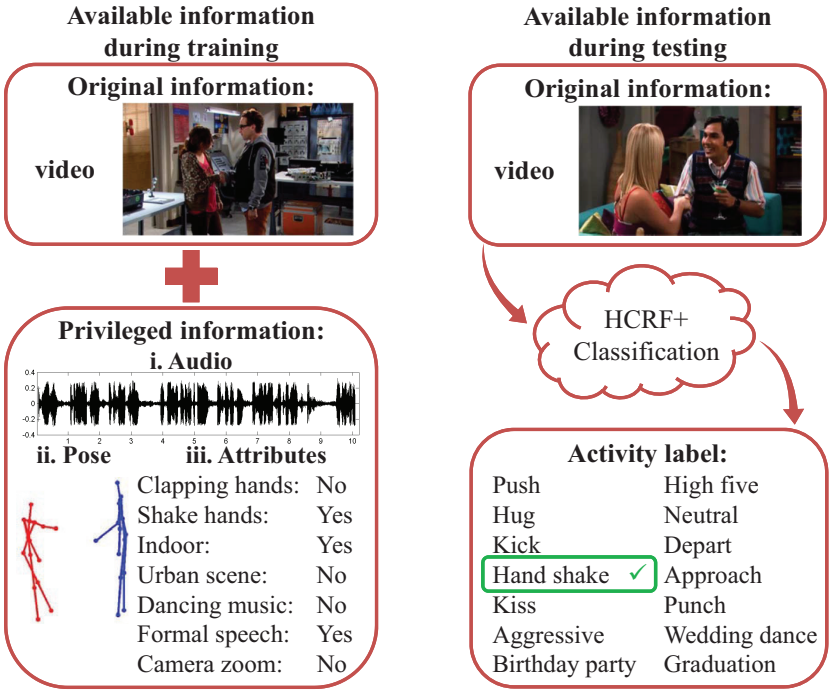

Fig. 1: Robust learning using privileged information. Given a set of training examples and a set of additional information about the training samples (left), our system can successfully recognize the class label of the underlying activity without having access to the additional information during testing (right). We explore three different forms of privileged information (e.g., audio signals, human poses, and attributes) by modeling them with a Student's t-distribution and incorporating them into the HCRF+ model.

The main contributions of our work can be summarized in the following points. A human activity recognition method is proposed, which exploits privileged information in a probabilistic manner by introducing a classification scheme based on HCRFs to deal with missing or incomplete data during testing. Both maximum likelihood and maximum margin approaches are incorporated into the proposed HCRF+ model. Moreover, a novel technique for adaptive estimation of the regularization term during the learning process is introduced by incorporating both privileged and regular data. Finally, contrary to previous methods, which may be sensitive to outlying data measurements, a robust framework for recognizing human activities is intergraded by employing a Student's t-distribution to attain robustness against outliers.

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 HCRF+ approach including the maximum likelihood and maximum margin approaches for learning the model's parameters and the automatic estimation of the regularization terms. In Section 4, experimental results are reported, and a discussion about the per-

formance of the proposed approach is offered in Section 5. Finally, conclusions are drawn in Section 6.

2 Related Work

A major family of methods relies on learning human activities by building visual models and assigning activity roles to people associated with an event [48,65]. In recent years, there has been an increased focus on the combination of different kinds of modalities for activity classification [64,13,15]. A shared representation of human poses and visual information has also been explored [73,4,37]. However, the effectiveness of such methods is limited by tracking inaccuracies in human poses and complex backgrounds. Unlike previous approaches, the work of Yan *et al.* [71] exploited spatio-temporal pose feature representations from three semantic pose modalities. Finally, a convolutional neural network (CNN) incorporates the learned pose representation as input to recognize human actions.

A special focus has also been given to recognizing human activities from movies or TV shows by exploiting scene contexts to localize activities and understand human interactions [41,20]. Ramanathan *et al.* [47] improved the recognition accuracy of such complex videos by relating textual descriptions and visual context to a unified framework. Guadarrama *et al.* [16] proposed an alternative to the previous approach that takes a video clip as input and generates short textual descriptions, which may correspond to an activity label that is unseen during training. However, natural video sequences may contain irrelevant scenes or scenes with multiple actions. Shao *et al.* [50] mixed appearance and motion features using multi-task deep learning for recognizing group activities in crowded scenes collected from the web. Marín-Jiménez *et al.* [38] used a bag of visual-audio words scheme along with late fusion for recognizing human interactions in TV shows. Even though their method performs well in recognizing human interaction, the lack of an intrinsic audio-visual relationship estimation limits the recognition problem.

Multiview human action recognition has also gained much popularity over the last decades. Gao *et al.* [14] proposed an adaptive fusion technique to combine information from multiple domains into a single processing pipeline. The authors constructed a category-level dictionary learning model to learn the adaptive weight of each camera and reweigh the learning samples according to their contribution to the action recognition task. In the same spirit, the work of Liu *et al.* [32] proposed a hierarchical dictionary learning-based method to encode local and global visual cues extracted from RGB and depth modalities for multiple-view human action recognition.

Intermediate semantic features representation for recognizing unseen actions during training has been extensively studied [33,12,75]. These intermediate features are learned during training and enable cross-stream fusion for capturing the correlations of true spatio-temporal features instead of treating appearance and motion features separately [9]. Also, the importance of visual relationship reasoning has been explored by Tsai *et al.* [56], where a gated fully-connected conditional random field was proposed to measure the relationship between different entities such as objects, attributes, subjects, or scenes, in the video sequence. Furthermore, a robust video-based human action recognition method that automatically infers the number of hidden states of a standard HCRF model directly from the input

data and coupled as a mixture of three Students t-components was also proposed by Vrigkas *et al.* [59].

Recent methods that exploited deep neural networks have demonstrated remarkable results in large-scale datasets [3,6]. Although large-scale datasets have proven to be a fundamental aspect of video action understanding, their inherent bias may lead to erroneous/controversial conclusions. Li *et al.* [31] minimized the representation bias by stitching together different action recognition datasets and extracting different levels of the representation hierarchy. The resulted dataset is an unbiased representation of the existing ones.

Perrett and Damen [44] proposed a cross-domain convolutional architecture, where long short-term memory (LSTM) networks [7] are used to learn temporal dependencies from two related to action recognition datasets. The LSTMs are connected to CNNs that can be jointly trained to simultaneously learn spatio-temporal dynamics. Wang *et al.* [66] presented a new video representation that employs CNNs to learn multi-scale convolutional feature maps and introduced the strategies of trajectory-constrained sampling and pooling to encode deep features into informative descriptors. Tran *et al.* [55] introduced a 3D ConvNet architecture that learns spatio-temporal features using 3D convolutions. Finally, a novel video representation, that can summarize a video into a single image by applying rank pooling on the raw image pixels, was proposed by Bilen *et al.* [1].

Feichtenhofer *et al.* [10] introduced a novel architecture for two-stream ConvNets and studied different ways for spatio-temporal fusion of the ConvNet towers. Zhu *et al.* [74] argued that videos contain one or more key volumes that are discriminative and most volumes are irrelevant to the recognition process. To this end, they proposed a unified deep learning framework to simultaneously identify discriminative key volumes and train classifiers, while they discarded all irrelevant volumes. On the contrary, Li *et al.* [30] argued that learning 2D rather than 3D convolutions can efficiently learn meaningful spatio-temporal features for video action-understanding.

The LUPI paradigm was first introduced by Vapnik and Vashist [57] 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+. Pechyony and Vapnik [42] formulated an algorithm for risk bound minimization with privileged information. Fouad *et al.* [11] proposed a combination of privileged information and metric learning. The privileged information was used to change the metric of the input data and thus any classifier could be used. Wand and Ji [70] also proposed two different loss functions that exploit privileged information and can be used with 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.

Wang *et al.* [69] incorporated privileged information in a latent max-margin model, where the additional knowledge was propagated through the latent nodes and the classification was performed from the regular data. Although this approach relaxes the strong assumptions of regular and privileged data relations for classification, it is limited by the slack variable estimation through SVM optimization. In this work, we address this problem by replacing the slack variables for the maximum margin violation and solve the unconstrained soft-margin SVM optimization problem. Smailis *et al.* [52] utilized the LUPI paradigm and the ResNet-34 model [19] to solve the problem of carrying human actions in still images and introduced

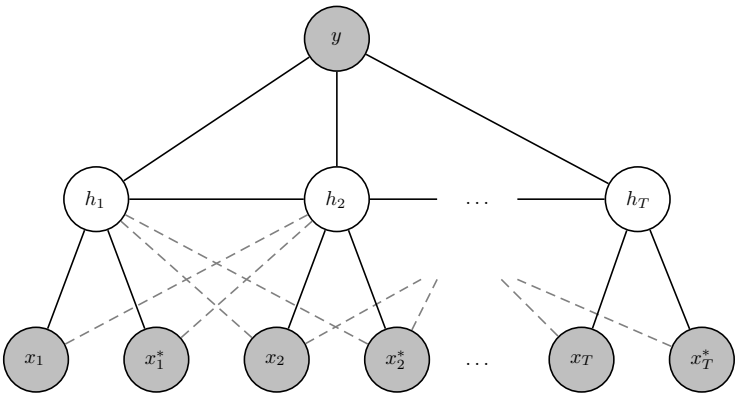

Fig. 2: Graphical representation of the chain structure model. The grey nodes are the observed features (x_i), the privileged information (x_i^*), and the unknown labels (y), respectively. The white nodes are the unobserved hidden variables (h).

a challenging dataset for carrying actions, which is formed by a few thousand images extracted from YouTube videos depicting several scenarios.

Serra-Toro *et al.* [49] 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 in [51]. Both regular and privileged features were considered of equivalent difficulty for recognizing the true class. Wang *et al.* [67] proposed a Bayesian network to learn the joint probability distribution of input features, output target, and privileged information. A combination of the LUPI framework and active learning has also been explored by Vrigkas *et al.* [62] to model human activities in a semi-supervised scheme. Recently, the LUPI paradigm has been employed with applications on gender classification, facial expression recognition, and hand pose estimation [23,63,72].

3 Methodology

Our method uses HCRFs, which are defined by a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 2), as the probabilistic framework for modeling the behavior of a subject in a video. During training, a classifier and the mapping from observations to the label set are learned. In testing, a probe sequence is classified into its respective state using loopy belief propagation (LBP) [27].

3.1 HCRF+ Model Formulation

We consider a labeled data set with N video sequences consisting of triplets $\mathcal{D} = \{(\mathbf{x}_{i,j}, \mathbf{x}^*_{i,j}, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_{i,j} \in \mathbb{R}^{M_{\mathbf{x}} \times T}$ is an observation sequence of length T with $j = 1 \dots T$. For example, $\mathbf{x}_{i,j}$ might correspond to the j^{th} frame of the i^{th} video sequence. Furthermore, y_i corresponds to a class label defined in a finite label

set $\mathcal{Y}$. In the context of robust learning using a privileged information paradigm, additional information about the observations $\mathbf{x}_i$ is encoded in a feature vector $\mathbf{x}^*_{i,j} \in \mathbb{R}^{M_{\mathbf{x}^*} \times T}$. Such privileged information is provided only at the training step and it is not available during testing. Note that we do not make any assumption about the form of the privileged data.

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

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

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

where $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters, and $\mathbf{h} = \{h_1, h_2, \ldots, h_T\}$, with $h_j \in \mathcal{H}$ being a set of latent variables. In particular, the number of latent variables may be different from the number of samples, as h_j may correspond to a substructure in an observation. Moreover, the features follow the structure of the graph, in which no feature may depend on more than two hidden states h_j and h_k [45]. This property not only captures the synchronization points between the different sets of information of the same state but also models the compatibility between pairs of consecutive states. We assume that our model follows the first-order Markov chain structure (i.e., the current state affects the next state). Finally, $E(y, \mathbf{h}|\mathbf{x}; \mathbf{w})$ is a vector of sufficient statistics and $A(\mathbf{w})$ is the log-partition function ensuring normalization:

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

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:

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

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned. Moreover, the potential functions correspond to the structure of the graphical model, as illustrated in Fig. 2. For example, a unary potential does not depend on more than two hidden variables h_j and h_k, and a pairwise potential may depend on h_j and h_k, which means that there must be an edge (j, k) in the graphical model.

The unary potential is expressed by:

$$\begin{aligned} \Phi(y, h_j, \mathbf{x}_j, \mathbf{x}^*_j; \boldsymbol{\theta}) = &\sum_{\ell} \phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) + \phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) \\ &+ \phi_3(h_j, \mathbf{x}^*_j; \boldsymbol{\theta}_3) , \end{aligned} \tag{4}$$

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

$$\phi_{1,\ell}(y, h_j; \boldsymbol{\theta}_{1,\ell}) = \sum_{\lambda \in \mathcal{Y}} \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_{1,\ell} \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a)\,, \tag{5}$$

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

$$\phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_2^\top \mathbb{1}(h_j = a)\mathbf{x}_j\,. \tag{6}$$

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

$$\phi_3(h_j, \mathbf{x}_j^*; \boldsymbol{\theta}_3) = \sum_{a \in \mathcal{H}} \boldsymbol{\theta}_3^\top \mathbb{1}(h_j = a)\mathbf{x}_j^*\,. \tag{7}$$

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

$$\Psi(y, h_j, h_k; \boldsymbol{\omega}) = \sum_{\substack{\lambda \in \mathcal{Y} \\ a,b \in \mathcal{H}}} \sum_{\ell} \boldsymbol{\omega}_\ell \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) \mathbb{1}(h_k = b)\,. \tag{8}$$

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

3.2 Maximum Likelihood Learning

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

$$L(\mathbf{w}) = \sum_{i=1}^{N} \frac{1}{\lambda_i} \log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \frac{1}{2\sigma^2} \|\mathbf{w}\|^2\,. \tag{9}$$

The first term is the log-likelihood of the posterior probability $p(y|\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ and quantifies how well the distribution in Eq. (1) defined by the parameter vector $\mathbf{w}$ matches the labels y, while λ is a tuning parameter. It can be rewritten as:

$$\begin{aligned} \log p(y_i | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) = &\log \sum_{\mathbf{h}} \exp(E(y, \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})) \\ &- \log \sum_{y' \neq y, \mathbf{h}} \exp(E(y', \mathbf{h} | \mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}))\,. \end{aligned} \tag{10}$$

The second term in Eq. (9) is a Gaussian prior with variance σ^2 and works as a regularizer. The use of hidden variables makes the optimization of the loss function non-convex, thus, a global solution is not guaranteed and we can estimate $\mathbf{w}^*$ that are locally optimal. The loss function in Eq. (9) is optimized using a gradient-descent method such as the limited-memory BFGS (LBFGS) method [39].

3.3 Maximum Margin Learning

We can easily transform the optimization problem of the loss function defined in Eq. (9) into a max-margin problem by substituting the log of the summation over the hidden states and the labels in Eq. (10) with maximization [68]. The goal is to maximize the margin between the score of the correct label and the score of the other labels. To learn the parameters $\mathbf{w}^*$ we need to minimize a loss function of the form:

$$\begin{aligned} L(\mathbf{w}) &= \sum_{i=1}^{N} \frac{1}{\lambda_i}\xi_i + \frac{1}{2\sigma^2}\|\mathbf{w}\|^2 \\ \text{s.t.} \max_{y'\neq y_i,\mathbf{h}} E(y', \mathbf{h}|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) &- \max_{\mathbf{h}} E(y_i, \mathbf{h}|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) \leq \xi_i - 1, \\ \text{and } \xi_i &\geq 0, \end{aligned} \tag{11}$$

where parameter λ is a tuning parameter. Although we add slack variables ξ to max-margin optimization, they eventually vanish. We do not estimate the slacks, but we replace them with the Hinge loss error [18] that penalizes the loss when the constraints in Eq. (11) are violated:

$$\ell_i(\mathbf{w}) = \max(0, 1 + (\max_{y'\neq y_i,\mathbf{h}} E(y', \mathbf{h}|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \max_{\mathbf{h}} E(y_i, \mathbf{h}|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}))) . \tag{12}$$

The optimization problem in (11) is equivalent to the optimization of the following unconstrained problem:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \frac{1}{\lambda_i}\ell_i(\mathbf{w}) + \frac{1}{2\sigma^2}\|\mathbf{w}\|^2 . \tag{13}$$

However, the quantity $\max(0, \cdot)$ is not differentiable and thus, Eq. (11) is hard to solve. To overcome this problem we adopt the bundle method of [54], which uses sub-gradient descent optimization algorithm.

3.4 Estimation of Regularization Parameters

Both maximum likelihood and max-margin loss functions introduce regularization parameters that control data fidelity and these regularization parameters in Eq. (9) and Eq. (13) may be obtained in closed form. Here, we examine the case of maximum likelihood optimization as the estimation of the regularization parameters for the max-margin optimization is equivalent. We can rewrite the loss function

in Eq. (9) as the sum of individual smoothing functionals for each of the training samples N:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \left\{ \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 \right\}, \tag{14}$$

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

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

$$\begin{aligned} \alpha_i(\mathbf{w}) &= f\left(\log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 \right) \\ &= \gamma_i \left\{ \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 \right\}, \end{aligned} \tag{15}$$

where γ is determined by the sufficient conditions for convergence. From Eq. (15), the regularization parameter α_i is computed as:

$$\alpha_i(\mathbf{w}) = \frac{\log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})}{\dfrac{1}{\gamma_i} + \|\mathbf{w}\|^2}, \tag{16}$$

and therefore:

$$\frac{1}{\gamma_i} > \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) - \alpha_i(\mathbf{w}) \|\mathbf{w}\|^2 . \tag{17}$$

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

$$\log p(y_i|\mathbf{x}_i; \mathbf{w}) \geq \log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w}) . \tag{18}$$

We can then relax the problem and consider that Eq. (17) is satisfied when $\frac{1}{\gamma_i} = \log p(y_i|\mathbf{x}_i; \mathbf{w})$. Thus, the regularization parameter α_i for the loss function is given by:

$$\alpha_i(\mathbf{w}) = \frac{\log p(y_i|\mathbf{x}_i, \mathbf{x}_i^*; \mathbf{w})}{\log p(y_i|\mathbf{x}_i; \mathbf{w}) + \|\mathbf{w}\|^2}, \tag{19}$$

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

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

$$\alpha_i(\mathbf{w}) = \frac{\ell_i(\mathbf{w})}{\zeta_i(\mathbf{w}) + \|\mathbf{w}\|^2}, \tag{20}$$

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

$$\zeta_i(\mathbf{w}) = \max(0, 1 + (\max_{y' \neq y_i, \mathbf{h}} E(y', \mathbf{h}|\mathbf{x}_i; \mathbf{w}) - \max_{\mathbf{h}} E(y_i, \mathbf{h}|\mathbf{x}_i; \mathbf{w}))) . \tag{21}$$

3.5 Inference

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

$$\begin{aligned} y &= \arg\max_y p(y|\mathbf{x};\mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y,\mathbf{h},\mathbf{x}^*|\mathbf{x};\mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y,\mathbf{h}|\mathbf{x},\mathbf{x}^*;\mathbf{w}) p(\mathbf{x}^*|\mathbf{x};\mathbf{w})\,. \end{aligned} \tag{22}$$

In the general case, the training samples $\mathbf{x}$ and $\mathbf{x}^*$ may be considered to be jointly Gaussian, thus the conditional distribution $p(\mathbf{x}^*|\mathbf{x};\mathbf{w})$ is also a Gaussian distribution. In the case of continuous features, the continuous space of features is quantized to a large number of discrete values to approximate the true value of the marginalization of Eq. (22). However, to efficiently cope with outlying measurements about the training data, we consider that the training samples $\mathbf{x}$ and $\mathbf{x}^*$ jointly follow a Student's t-distribution. Therefore, the conditional distribution $p(\mathbf{x}^*|\mathbf{x};\mathbf{w})$ is also a Student's t-distribution $\mathrm{St}(\mathbf{x}^*|\mathbf{x};\mu^*,\Sigma^*,\nu^*)$, where $\mathbf{x}^*$ forms the first $M_{\mathbf{x}^*}$ components of $(\mathbf{x}^*,\mathbf{x})^T$, $\mathbf{x}$ comprises the remaining $M - M_{\mathbf{x}^*}$ components, μ^* is the mean vector, Σ^* is the covariance matrix and $\nu^* \in [0,\infty)$ corresponds to the degrees of freedom of the distribution [28]. Note that by letting the degrees of freedom ν^* go to infinity, we can recover the Gaussian distribution with the same parameters. If the data contain outliers, the degrees of freedom parameter ν^* is weak and the mean and covariance of the data are appropriately weighted in order not to take into account the outliers. More details on how the parameters of the conditional Student's t-distribution $p(\mathbf{x}^*|\mathbf{x};\mathbf{w})$ are estimated can be found in Appendix A.

Although both conditional distributions $p(y,\mathbf{h}|\mathbf{x},\mathbf{x}^*;\mathbf{w})$ and $p(\mathbf{x}^*|\mathbf{x};\mathbf{w})$ belong to the exponential family, the graph in Fig. 2 is cyclic, and therefore an exact solution to Eq. (22) is generally intractable. For this reason, approximate inference is employed for estimation of the marginal probability by applying the LBP algorithm [27].

4 Experimental Results

We evaluated our method on four challenging publicly available datasets. Three different types of privileged information were used: audio signal, human pose, and semantic attribute annotation.

We propose four variants of our approach, called *Maximum Likelihood LUPI Hidden Conditional Random Field (ml-HCRF+)*, *Adaptive Maximum Likelihood LUPI Hidden Conditional Random Field (aml-HCRF+)*, *Maximum Margin LUPI Hidden Conditional Random Field (mm-HCRF+)*, and *Adaptive Maximum Margin LUPI Hidden Conditional Random Field (amm-HCRF+)*, depending on which learning method we apply (i.e., maximum likelihood or max-margin) and whether

we automatically estimate the regularization parameters of the corresponding loss function or not.

4.1 Datasets

Parliament [60]: This dataset is a collection of 228 video sequences, depicting political speeches in the Greek parliament, at a resolution of 320×240 pixels at 25 fps. The video sequences were manually labeled with one of three behavioral labels: *friendly*, *aggressive*, or *neutral*.

TV human interaction (TVHI) [41]: This dataset consists of 300 video sequences collected from over 20 different TV shows. The video clips contain four kinds of interactions: handshakes, high fives, hugs, and kisses, equally split into 50 video sequences each, while the remaining 100 video clips do not contain any of the aforementioned interactions.

SBU Kinect Interaction (SBU) [73]: This dataset contains approximately 300 video sequences depicting two-person interactions captured by a Microsoft Kinect sensor. The dataset contains eight different classes including approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands, which are performed by seven different persons. It also contains three-dimensional coordinates of 15 joints for each person at each frame.

Unstructured social activity attribute (USAA) [12]: The USAA dataset includes eight different semantic class videos of social occasions such as birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance, and wedding reception. It contains around 100 videos per class for training and testing. Each video is annotated with 69 attributes, which can be divided into five broad classes: actions, objects, scenes, sounds, and camera movement.

4.2 Implementation Details

Deep learning model: In our experiments, we used CNNs for both end-to-end classification and feature extraction. We employed the pre-trained model of Tran *et al.* [55], which is a 3D ConvNet (C3D) as it can be seen in Fig. 3. We selected this model because it was trained on a very large dataset (Sports 1M [24]), which provides good features for the activity recognition task, especially in our case where the size of the training data is small, making deep learning models prone to overfitting.

Because both the Parliament and SBU datasets are fairly small datasets, only a few parameters had to be trained to avoid overfitting. Particularly, we replaced the fully-connected layers of the pre-trained model with a new fully-connected layer of size $1,024$ and trained the additional layer coupled with a softmax layer on top of it. For the TVHI dataset, we fine-tuned the last group of convolutional layers, while for the USAA dataset, we fine-tuned the last two groups. Each group has two convolutional layers, while we added a new fully-connected layer of size 256 for the TVHI and $1,024$ for the USAA datasets, respectively. For the optimization process, we used mini-batch stochastic gradient descent (SGD) with momentum. The size of the mini-batch was set to 16 and we used a constant momentum of 0.9.

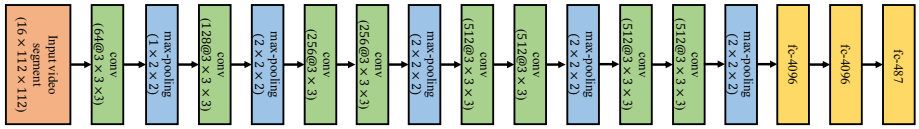

Fig. 3: Illustration of the 3D ConvNet architecture [55]. The model has 8 convolutional layers, 5 max-pooling layers, and two fully-connected layers followed by the output layer, which is a softmax classifier that classifies videos in 487 categories [24]. Different colors are used for different types of layers for better understanding. The net takes as input a volume of 16 frames with a height and a width of size 112. Convolutional layers perform 3D convolutions using 3D kernels and max-pooling layers perform 3D pooling using 3D receptive fields. In convolutional layers, the number of feature maps is denoted before symbol "@" and follows the size of the kernels. In pooling layers, the size of each receptive field is shown. In fully-connected layers, it is denoted the number of hidden units.

For both the Parliament and SBU datasets, the learning rate was initialized to 0.01 and it was decayed by a factor of 0.1, while the total number of training epochs was $1,000$. For the TVHI and USAA datasets, we used a constant learning rate of 10^{-4}, and the total number of training epochs was 500 and 250, respectively. For all datasets, we added a dropout layer after the new fully-connected layer with a probability of 0.5. Also, we performed data augmentation on each batch online and 16 consecutive frames were randomly selected for each video. These frames were randomly cropped, resulting in frames of size 112×112 and then flipped with probability 0.5. For the classification task, we used the centered 112×112 crop on the frames of each video sequence. Then, for each video, we extracted 10 random clips of 16 frames and averaged their predictions. Finally, to avoid overfitting, we used early stopping and extracted C3D features from the newly added fully-connected layer.

Privileged Information: For our experiments on Parliament and TVHI datasets, we used audio features as privileged information. More, specifically, we employed the mel-frequency cepstral coefficients (MFCC) [46] features and their first and second-order derivatives. The audio signal was sampled at 16 KHz and processed over 10 *ms* using a Hamming window with a 25% overlap. The audio feature vector consisted of a collection of 13 MFCC coefficients along with the first and second derivatives forming a 39 dimensional audio feature vector. Furthermore, for the SBU dataset, privileged information is represented by a 15-dimensional feature vector capturing human pose information and information from the joints that correspond to joint distance, joint motion, plane, normal plane, velocity, and normal velocity as described in [73]. Finally, for the USAA dataset, we used the provided attribute annotation as privileged information to characterize each class not with an individual label, but with a feature vector of semantic attributes.

Hand-crafted feature selection: For the evaluation of our method, we used spatio-temporal interest points (STIP) [29] as our base video representation. First, we extracted local space-time features at a rate of 25 fps using a 72-dimensional vector of HoG and 90-dimensional vector of HoF feature descriptors [26] for each STIP, which captures the human motion between frames. These features were

selected because they can efficiently and compactly capture salient visual motion patterns. Besides, for the TVHI dataset, we also used the provided annotations, which are related to the locations of the persons in each video clip, including the head orientations of each subject in the clips, the pair of subjects who interact with each other, and the corresponding labels. Moreover, for the USAA dataset as a representation of the video data, we used the provided low-level features, which correspond to SIFT [35], STIP, and MFCC features.

Model selection: The proposed model was trained by varying the number of hidden states from 3 to 20, with a maximum of 400 iterations for the termination of the LBFGS optimization method. The L_2 regularization scale term σ for the non-adaptive methods was set to 10^k, with $k \in \{-3, \ldots, 3\}$. The evaluation of our method was performed using 5-fold cross-validation to split the datasets into training and test sets, and the average results over all the examined configurations are reported.

4.3 Multimodal Feature Fusion

One drawback of combining features of different modalities is the different probability distribution that each modality may have. Thus, instead of directly combining multimodal features, one may employ canonical correlation analysis (CCA) [17] to exploit the correlation between the different modalities by projecting them onto a common subspace so that the correlation between the input vectors is maximized in the projected space. In this paper, we followed a different approach. Our model can learn the relationship between the input data and the privileged features. To this end, we jointly calibrate the different modalities by learning a multiple output linear regression model [40]. Let $\mathbf{x} \in \mathbb{R}^{M \times d}$ be the input raw data and $\mathbf{x}^* \in \mathbb{R}^{M \times p}$ be the set of privileged features. Our goal is to find a set of weights $\boldsymbol{\gamma} \in \mathbb{R}^{d \times p}$, which relates the privileged features to the regular features by minimizing a distance function across the input samples and their attributes:

$$\arg\min_{\boldsymbol{\gamma}} \|\mathbf{x}\boldsymbol{\gamma} - \mathbf{x}^*\|^2 + \eta\|\boldsymbol{\gamma}\|^2, \tag{23}$$

where $\|\boldsymbol{\gamma}\|^2$ is a regularization term and η controls the degree of the regularization, which was chosen to give the best solution by using cross validation with $\eta \in [10^{-4}, 1]$. Following a constrained least squares (CLS) optimization problem and minimizing $\|\boldsymbol{\gamma}\|^2$ subject to $\mathbf{x}\boldsymbol{\gamma} = \mathbf{x}^*$, Eq. (23) has a closed form solution $\boldsymbol{\gamma} = \left(\mathbf{x}^T\mathbf{x} + \eta I\right)^{-1}\mathbf{x}^T\mathbf{x}^*$, where I is the identity matrix. Note that the minimization of Eq. (23) is fast since it needs to be solved only once during training. Finally, we obtain the prediction f of the privileged features by multiplying the regular features with the learned weights $f = \mathbf{x} \cdot \boldsymbol{\gamma}$. The main steps of the proposed method are summarized in Algorithm 1.

4.4 Comparisons using Deep Learning Features

In this section, we compared the proposed HCRF+ method with the LSTM networks [21], since it has been proven that they provide good performance in several sequential classification tasks such as image description and activity recognition

Algorithm 1: Robust privileged probabilistic learning

Input : Training sets $\mathcal{X}$, and $\mathcal{X}^*$, training labels $\mathcal{Y}$
Perform feature extraction from both $\mathcal{X}$ and $\mathcal{X}^*$
Employ Eq. (23) and project $\mathcal{X}$ and $\mathcal{X}^*$ onto a common space
Initialize parameters $\mathbf{w}$ randomly
for $i \in \{1, \ldots, N\}$ **do**
  /*Maximum likelihood or max-margin learning*/ Estimate the regularization parameter α_i using Eqs. (19) or (20)
  $\mathbf{w}^* \leftarrow$ Train HCRF+ on triplets $(\mathcal{X}_i, \mathcal{X}_i^*, \mathcal{Y}_i)$
end
Output: Estimated models' parameters $\mathbf{w}^*$

Table 1: Comparison of the classification accuracies (%) on Parliamment [60], TVHI [41], SBU [73], and USAA [12] datasets using C3D features. Results highlighted with light purple indicate statistically significant improvement using paired t-test.

Method	Parliament	TVHI	SBU	USAA
Methods without privileged information				
HCRF [45]	84.4 ± 0.8	89.6 ± 0.5	91.1 ± 0.4	91.6 ± 0.8
SVM [2]	89.9 ± 0.5	90.0 ± 0.3	92.8 ± 0.2	91.9 ± 0.3
3D ConvNet (C3D) [55]	78.1 ± 0.4	60.5 ± 1.1	94.2 ± 0.8	67.4 ± 0.6
LSTM [21]	88.3 ± 0.8	88.4 ± 1.5	94.7 ± 0.7	91.3 ± 1.7
Methods with privileged information				
SVM+ [57]	90.0 ± 0.3	92.5 ± 0.4	94.8 ± 0.3	92.3 ± 0.3
Wang and Ji [70]	83.5 ± 0.4	88.8 ± 0.2	92.7 ± 0.4	92.8 ± 0.2
Wang *et al.* [69]	84.4 ± 0.6	85.0 ± 1.2	91.1 ± 1.3	93.2 ± 1.2
Sharmanska *et al.* [51]	81.8 ± 0.2	90.0 ± 0.1	92.9 ± 0.4	93.5 ± 0.2
ml-HCRF+	$\mathbf{93.3} \pm 0.7$	$\mathbf{93.2} \pm 0.6$	$\mathbf{94.9} \pm 0.7$	93.9 ± 0.9
aml-HCRF+	$\mathbf{93.3} \pm 0.4$	92.5 ± 1.1	92.9 ± 0.4	95.9 ± 1.3
mm-HCRF+	88.9 ± 0.9	92.5 ± 0.7	93.6 ± 1.1	95.2 ± 1.0
amm-HCRF+	86.7 ± 1.2	90.0 ± 0.8	94.6 ± 1.0	$\mathbf{96.4} \pm 1.4$

[8]. Although a promising methodology is to train a CNN stacked with an LSTM layer on top [8] for end-to-end feature extraction and sequential classification, our limited size datasets prevented us from training such a model due to overfitting. To address this issue, we trained an LSTM layer with a softmax layer on top, of the features extracted from the pre-trained CNN model. Specifically, we added a dropout layer on the LSTM's hidden units and an L_2 regularization on the softmax units. For the estimation of the hyperparameters, we performed a grid search with 5-fold cross-validation to optimize the learning rate, the number of hidden units, the dropout rate, and the weight decay factor of the L_2 regularizer. We trained the LSTM model for 100 epochs using the Adam optimizer [25] with early stopping.

The comparison of the proposed approach with state-of-the-art methods using the C3D features is summarized in Table 1. In particular, to show the benefit of using robust privileged information, we compared our method both with state-of-the-art methods with and without incorporating the LUPI paradigm. Also,

to demonstrate the efficacy of the robust privileged information to the problem of human activity recognition, we compared it with ordinary SVM and HCRF, as if they could access both the regular 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. Furthermore, for the SVM+ and SVM, we consider a one-versus-one decomposition of a multi-class classification scheme and average the results for every possible configuration. Finally, the optimal parameters for the SVM and SVM+ were selected using cross-validation.

It is worth noting that privileged information works in favor of the classification task in all cases. The ml-HCRF+ variant achieves the highest results among all other methods for the Parliament, TVHI, and SBU datasets, while for the USAA dataset, the amm-HCRF+ variant achieves the highest recognition accuracy (96.4%). Moreover, the improvement in the accuracy of the proposed model concerning the C3D classification for the Parliament, TVHI, and USAA datasets, was approximately 15%, 33%, and 29%, respectively. This improvement can be explained by the fact that the C3D model uses a linear classifier in the softmax layer, while the proposed approach is a more sophisticated model that can efficiently handle sequential data in a more principled way. Also, the performance improvement, brought by the LSTM compared to the 3D ConvNet, validates the ability of LSTMs to capture long-term dependencies in human activities as LSTMs have a memory of previous activity states and can better model their complex dynamics. Nonetheless, the proposed model outperforms the LSTM, for all datasets, a fact that supports our main hypothesis that the LUPI paradigm may be beneficial for human activity recognition.

The corresponding confusion matrices of the proposed method for all datasets, using the C3D features, are depicted in Fig. 4. The combination of privileged information with the feature representation learned from the C3D model resulted in very small inter- and intra-class classification errors for all datasets. For the SBU dataset, only a few classes are confused with each other (e.g., the class *kick* versus the class *push*), while four out of the eight classes were perfectly recognized.

A comparison of the mean per-class accuracies of the proposed approach with state-of-the-art methods on all four datasets is illustrated in Fig. 5. On the Parliament dataset, the proposed ml-HCRF+ method has the highest recognition accuracy (97.6%) among the other variants of the proposed model, while it achieves the same accuracy as the standard HCRF model. It is also worth mentioning that our method can increase the recognition accuracy by nearly 38% concerning the methods of Wang and Ji [70] and the method of Sharmanska et al. [51], which also incorporate the LUPI paradigm. This significantly high increase in recognition accuracy indicates the strength of the proposed method. For the TVHI dataset, we significantly managed to increase the classification accuracy by approximately 10%, concerning the LUPI-based SVM+ and Wang and Ji [70] approaches, as our approach achieves very high recognition accuracy (84.9%). The improvement of our method compared to the method of Sharmanska et al. [51] and the methods that do not use privileged information was even higher. In Fig. 5(c), the ml-HCRF+ approach achieved the highest accuracy (85.4%), where the improvement over the standard HCRF model is nearly 4%. Comparing our method to methods that do not use privileged information, we increased the classification accuracy in all cases. For the USAA dataset, the combination of both raw data and attribute represen-

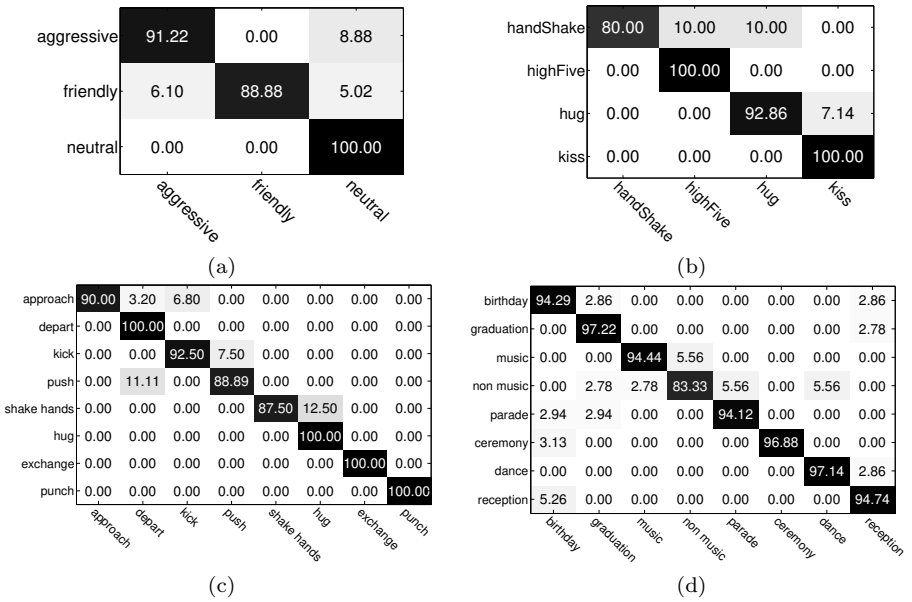

Fig. 4: Confusion matrices of the proposed ml-HCRF+ approach for (a) the Parliament [60], (b) the TVHI [41], (c) the SBU [73], and (d) the USAA [12] datasets using the CNN features.

tation of human activities significantly outperformed the SVM+ baseline and the method of Wang and Ji [70] by increasing the classification accuracy by approximately 11% for the amm-HCFR+ model. An improvement of 3% concerning the methods of Sharmanska et al. [51] and Wang et al. [69] was also achieved.

Although the adaptive HCRF+ approaches may perform worse than the non-adaptive variants, they can still achieve better results than the majority of the state-of-the-art methods. One reason for this is that the estimation of the regularization parameters for the adaptive variants depends on the input features. Features that belong to the background may influence the estimation of the regularization parameters as they may serve as background noise. Also, the Parliament dataset contains large intra-class variabilities. For example, the interaction between an arm lift and the raise in the voice may not exclusively be combined together as some features may act as outliers and affect the classification accuracy. However, it is interesting to observe that, for the USAA dataset, the adaptive variants of the proposed method perform better than their non-adaptive counterparts. Automatic estimation of the regularization parameters provides more flexibility to the model as it allows the model to adjust its behavior according to the training data.

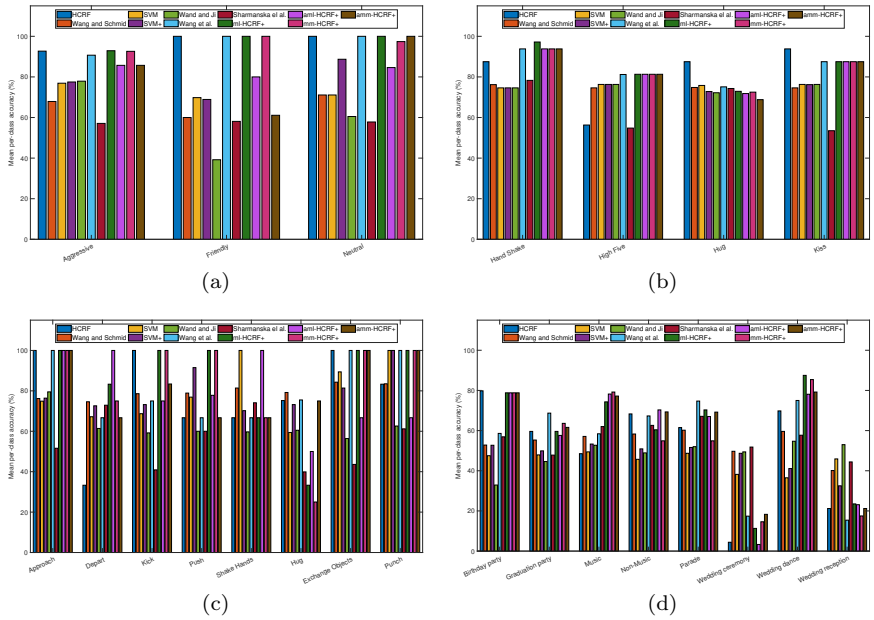

Fig. 5: Mean per-class accuracies showing the performance of the proposed method with respect to the state of the art for (a) the Parliament [60], (b) the TVHI [41], (c) the SBU [73], and (d) the USAA [12] datasets.

4.5 Comparisons using Hand-Crafted Features

A comparison of the proposed approach with state-of-the-art methods using hand-crafted features is depicted in Table 2. The proposed ml-HCRF+ method has the highest recognition accuracy among the other variants of the proposed model.

Also, note that the improvement of accuracy using the C3D features (Table 1 with respect to the hand-crafted feature classification (Table 2), for all datasets, indicates that CNNs may efficiently extract informative features without any need to hand-design them.

The resulting confusion matrices of the best performing variant are depicted in Fig. 6. It is worth mentioning that for both the Parliament and the TVHI datasets the classification errors between different classes are relatively small. It is interesting to observe that for the USAA dataset the different classes may be strongly confused. For example, the class *wedding ceremony* is confused with the class *graduation party* and the class *wedding reception* is confused with the class *non-music performance*. This is because the different classes may share the same attribute representation as different videos may have been captured under similar conditions.

Table 2: Comparison of the classification accuracies (%) on Parliamment [60], TVHI [41], SBU [73], and USAA [12] datasets using had-crafted features. Results highlighted with light purple indicate statistically significant improvement over the second best method using paired t-test.

Method	Parliament	TVHI	SBU	USAA
Methods without privileged information				
HCRF	**97.6** $\pm$ 0.6	81.3 $\pm$ 0.7	81.4 $\pm$ 0.8	54.0 $\pm$ 0.8
Wang and Schmid [65]	66.6 $\pm$ 0.5	76.1 $\pm$ 0.4	79.6 $\pm$ 0.4	55.6 $\pm$ 0.1
SVM [2]	72.6 $\pm$ 0.4	75.9 $\pm$ 0.6	79.4 $\pm$ 0.4	47.4 $\pm$ 0.1
Methods with privileged information				
SVM+ [57]	78.4 $\pm$ 0.2	75.0 $\pm$ 0.2	79.4 $\pm$ 0.3	48.5 $\pm$ 0.1
Wang and Ji [70]	59.2 $\pm$ 0.2	74.8 $\pm$ 0.2	62.4 $\pm$ 0.3	48.5 $\pm$ 0.2
Wang *et al.* [69]	96.9 $\pm$ 1.1	84.4 $\pm$ 1.1	83.7 $\pm$ 1.6	55.3 $\pm$ 0.9
Sharmanska *et al.*[51]	57.7 $\pm$ 0.4	65.2 $\pm$ 0.1	56.3 $\pm$ 0.2	56.3 $\pm$ 0.2
ml-HCRF+	**97.6** $\pm$ 0.7	**84.9** $\pm$ 0.8	**85.4** $\pm$ 0.4	58.1 $\pm$ 1.4
aml-HCRF+	83.5 $\pm$ 1.3	83.6 $\pm$ 1.1	79.8 $\pm$ 1.3	57.5 $\pm$ 1.4
mm-HCRF+	96.5 $\pm$ 0.9	83.6 $\pm$ 0.6	83.7 $\pm$ 0.5	56.8 $\pm$ 0.6
amm-HCRF+	82.3 $\pm$ 1.3	82.9 $\pm$ 0.8	82.8 $\pm$ 1.3	**59.4** $\pm$ 0.7

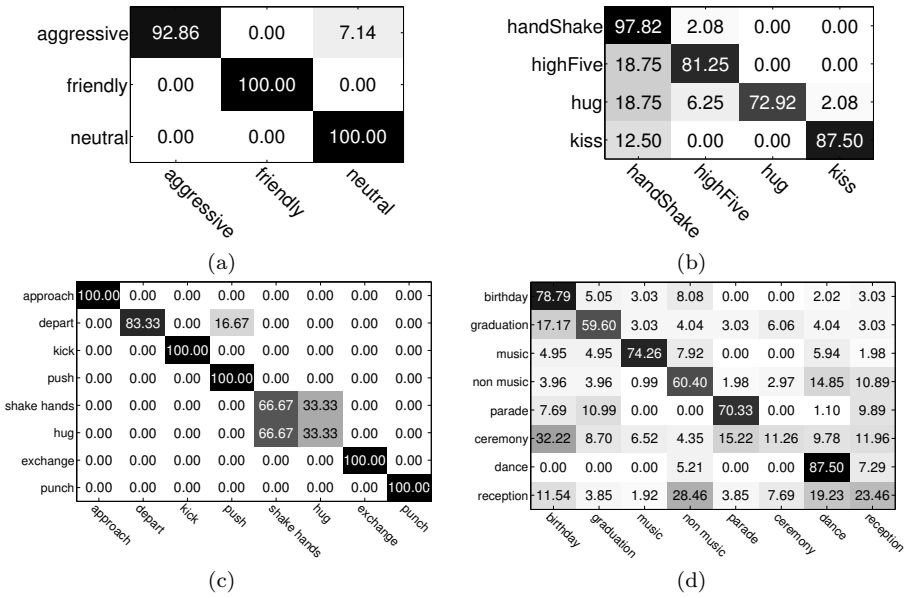

Fig. 6: Confusion matrices of the proposed ml-HCRF+ approach for (a) the Parliament [60], (b) the TVHI [41], (c) the SBU [73], and (d) the USAA [12] datasets.

Table 3: Ablation studies to assess the impact of each type of information on the final result using the standard HCRF model. The checkmark corresponds to the usage of specific information for training and testing.

Dataset	Regular	Privileged	Accuracy
Parliament [60]	✓		67.1 ± 1.4
		✓	72.7 ± 1.8
	✓	✓	97.6 ± 0.7
TVHI [41]	✓		60.9 ± 1.3
		✓	35.9 ± 1.5
	✓	✓	81.3 ± 0.7
SBU [73]	✓		69.8 ± 1.1
		✓	62.5 ± 1.3
	✓	✓	81.4 ± 0.8
USAA [12]	✓		55.5 ± 0.9
		✓	37.4 ± 1.0
	✓	✓	54.0 ± 0.8

4.6 Ablation Studies

To complete the study, we also trained an HCRF model that uses only the regular or only the privileged information for both training and testing. In Table 3, we investigate to what extent each type of information affects the final performance. If only privileged information is used as regular features for classification, the recognition accuracy is notably lower than when using only the regular information for the classification task. In general, when privileged information alone is used as regular information may not be enough for the correct classification of an action label into its respective category. This is because it is commonplace that finding proper privileged information is not always a straightforward problem.

To our surprise, we observed that, for the SBU dataset, even though for some classes we were able to perfectly recognize the underlying activity, the model failed though to recognize some of the classes as the rate of false positives may reach 100%. This reinforces the fact that different modalities may help in constructing better classifiers.

Note that for the USAA dataset, when privileged information is used as regular, the performance of the model drops in terms of accuracy by 18%. We believe that this is because the use of binary features for training and testing may cause bias. However, the combination of regular and privileged information durion training and testing does not suffer from the biasing problem due to feature calibration and their projection to a common subspace using Eq. (23).

The classification accuracy to the number of hidden states is depicted in Fig. 7. We may observe that all four variants have a similar behavior as the number of hidden states increases. It is clear that when privileged information is used, in the vast majority of the cases (38 out of 45 cases) all variants of HCRF+ perform better than the standard HCRF model. In Fig. 7, the HCRF+ variants and the standard HCRF model suffer from large fluctuations as the number of hidden states increases. This is because the number of hidden states plays a crucial role

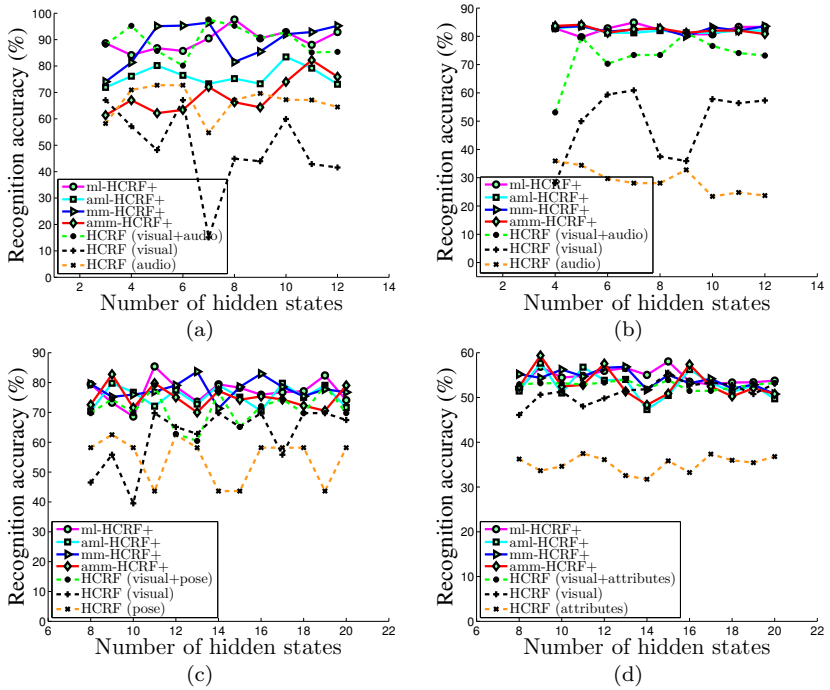

Fig. 7: Comparison of the recognition accuracy of the four different variants of the proposed method and standard HCRF model with respect to the number of hidden states for (a) the Parliament [60], (b) the TVHI [41], (c) the SBU [73], and (d) the USAA [12] datasets. The text in parentheses in the legend of each figure corresponds to the type of information used both for training and testing.

in the recognition process. Many hidden states may lead to model overfitting, while few hidden states may cause underfitting. This would be resolved by the estimation of the optimal number of hidden states during learning, but this is not straightforward for this model. We may also observe that the performance of each modality alone is kept significantly lower for all configurations of hidden states, which reinforces the fact that privileged information may help to construct better classification models.

The behavior of the proposed adaptive model as a function of the regularization parameters and the number of hidden states is depicted in Fig. 8. To be consistent to the non-adaptive methods, the real-valued regularization parameters were quantized from the continuous to the discrete space with $\alpha(\mathbf{w}) = 10^k, k \in \{-2, \ldots, 2\}$ and the results were averaged. We may observe that the behavior of the recognition accuracy is smooth for the different values of $\alpha(\mathbf{w})$ and the number of hidden states, which indicates that the automatic estimation of $\alpha(\mathbf{w})$ is robust.

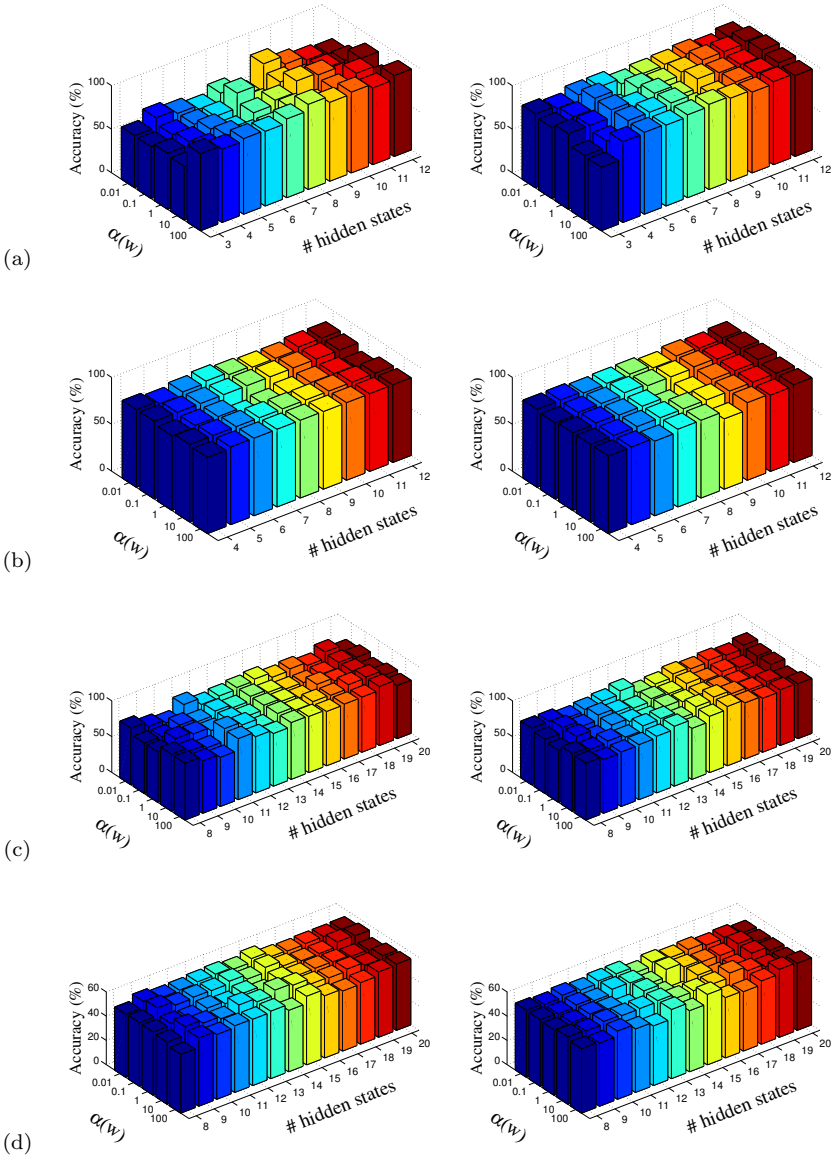

Fig. 8: Recognition performance of the proposed maximum likelihood (left-column) and max-margin (right-column) variants as function of the regularization parameter and the number of hidden states for (a) the Parliament [60], (b) the TVHI [41], (c) the SBU [73] and (d) the USAA [12] datasets.

5 Discussion

Our method can robustly use privileged information in a more efficient way than the SVM+ and the other LUPI-based methods, by exploiting the hidden dynamics between the video clips and the privileged information. We may also observe that the proposed method outperforms all methods that do not incorporate privileged information during learning. Since the combination of multimodal data falls natural to the human perception of understanding complex activities, the incorporation of such information constitutes a strong attribute for discriminating between different classes, rather than learning each modality separately.

Statistical significance tests: To provide statistical evidence of the recognition accuracy, we computed the p-values of the obtained results to the compared methods. Results highlighted with light purple in Table 1 and Table 2 indicate statistically significant improvement (p-values were less than the significance level of 0.05) over the second-best method using paired t-test. In general, we may conclude that the improvements obtained by our model are statistically significant and not due to chance.

Computational complexity: The proposed method uses the same sufficient statistics as HCRF and the computational complexity is similar to HCRF. The complexity of our method is determined by the complexity of the corresponding inference problem and is quadratic to the number of hidden states.

5.1 Why is Privileged Information Important?

Selecting which features can act as privileged information is not an easy task. The performance of LUPI-based classifiers relies on the delicate relationship between the regular and the privileged information. Also, privileged information is costly or difficult to obtain with respect to producing additional regular training examples [49]. In general, when privileged information alone is used as regular it may not be sufficient for the correct classification of an action label into its respective category, since finding proper privileged information is not always a straightforward process.

The scope of our approach is not to achieve the best results possible but to investigate to what extent privileged information can be beneficial under the same evaluation protocol. The main strength of the proposed method is that it achieves good classification results when the LUPI framework is incorporated with the standard HCRF model.

In the era of deep learning, significant progress has been made in learning good representations of the data and a deep learning-based technique is a way to go. However, in cases where datasets are small in size, which is true in our case, and the distribution of the data is completely different from the data that the existing pre-trained models were trained on, then privileged information can be very helpful. Nonetheless, one may fine-tune the deep neural model and extract meaningful feature representations. This enhances our choice to use deep features with the proposed HCRF+ model as the experimental results indicate significant improvement when these features are used. Thus, the answer to the question "is privileged information necessary?" is affirmative. For example, in many medical applications, where pre-trained deep learning models are still not available, privileged information is the best solution to go.

6 Conclusion

In this paper, we addressed the problem of human activity categorization in a supervised framework and proposed a novel probabilistic classification model based on robust learning using a privileged information paradigm, called HCRF+. Our model is made robust using Student's t-distributions to model the conditional distribution of the privileged information. We proposed two variants for training in the LUPI framework. The first variant uses maximum likelihood and the second uses maximum margin learning.

Using auxiliary information about the input data, we were able to produce better classification results than the standard HCRF [45] approach. We evaluated the performance of our method on four publicly available datasets and tested various forms of privileged information. The experimental results indicated that robust privileged information along with the regular input data for training the model ameliorates the recognition performance. We demonstrated improved results concerning the state-of-the-art LUPI framework especially when C3D features are employed.

According to our results, the proposed method and its variants achieved notably higher performance than the majority of the compared classification schemes. We were able to flexibly understand multimodal human activities with high accuracy when not the same amount of information is available during testing. By automatically estimating the regularization parameters during learning, we managed to achieve high recognition accuracy with less effort than standard cross-validation based classification schemes.

A Conditional Distribution of the Privileged Information

Recall that $\mathbf{x} \in \mathbb{R}^{M_{\mathbf{x}} \times T}$ is an observation sequence of length T and $\mathbf{x}^* \in \mathbb{R}^{M_{\mathbf{x}^*} \times T}$ corresponds to the privileged information of the same length. We partition the original set $(\mathbf{x}^*, \mathbf{x})^T \in \mathbb{R}^{M \times T}$ into two disjoint subsets, where $\mathbf{x}^*$ forms the first $M_{\mathbf{x}^*}$ components of $(\mathbf{x}^*, \mathbf{x})^T \in \mathbb{R}^{M \times T}$ and $\mathbf{x}$ comprises the remaining $M - M_{\mathbf{x}}$ components. If the joint distribution $p(\mathbf{x}, \mathbf{x}^*; \mathbf{w})$ follows a Student's t-law, with mean vector $\mu = (\mu_{\mathbf{x}^*}, \mu_{\mathbf{x}})^T$, a real, positive definite, and symmetric $M \times M$ covariance matrix $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{x}^*\mathbf{x}^*} & \Sigma_{\mathbf{x}^*\mathbf{x}} \\ \Sigma_{\mathbf{x}\mathbf{x}^*} & \Sigma_{\mathbf{x}\mathbf{x}} \end{pmatrix}$ and $\nu \in [0, \infty)$ corresponds to the degrees of freedom of the distribution [28], then the conditional distribution $p(\mathbf{x}|\mathbf{x}^*; \mathbf{w})$ is also a Student's t-distribution:

$$
\begin{aligned}
p(\mathbf{x}^*|\mathbf{x}; \mathbf{w}) &= \mathrm{St}(\mathbf{x}^*; \mu^*, \Sigma^*, \nu^*) \\
&= \frac{\Gamma\left(\left(\nu^* + M\right)/2\right) |\Sigma_{\mathbf{x}\mathbf{x}}|^{1/2}}{\left(\pi\nu^*\right)^{M_{\mathbf{x}}/2} \Gamma\left(\left(\nu^* + M_{\mathbf{x}}\right)/2\right) |\Sigma^*|^{1/2}} \\
&\quad \times \frac{\left[1 + \frac{1}{\nu^*} \mathbf{x}^T \Sigma_{\mathbf{x}\mathbf{x}}^{-1} \mathbf{x}\right]^{\frac{(\nu^* + M_{\mathbf{x}})}{2}}}{\left[1 + \frac{1}{\nu^*} Z^T \Sigma^{*-1} Z\right]^{\frac{(\nu^* + M)}{2}}} .
\end{aligned}
\tag{24}
$$

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

$$\mu^* = \mu_{\mathbf{x}^*} - \Sigma_{\mathbf{x}^*\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\left(\mathbf{x}-\mu_{\mathbf{x}}\right) , \tag{25}$$

$$\Sigma^* = \frac{\nu_{\mathbf{x}^*} + \left(\mathbf{x}-\mu_{\mathbf{x}}\right)^T \Sigma_{\mathbf{xx}}^{-1}\left(\mathbf{x}-\mu_{\mathbf{x}}\right)}{\nu_{\mathbf{x}^*} + M_{\mathbf{x}^*}} \times \left(\Sigma_{\mathbf{x}^*\mathbf{x}^*} - \Sigma_{\mathbf{x}^*\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xx}^*}\right) , \tag{26}$$

$$\nu^* = \nu_{\mathbf{x}^*} + M_{\mathbf{x}^*} . \tag{27}$$

The parameters (μ, Σ, ν) of the joint Student's t-distribution $p(\mathbf{x}^*, \mathbf{x}; \mathbf{w})$, which are defined by the corresponding partition of the vector $(\mathbf{x}^*, \mathbf{x})^T$, are estimated using the expectation-maximization (EM) algorithm [28]. Then, the parameters of the conditional distribution $p(\mathbf{x}^*|\mathbf{x};\mathbf{w})$ are computed using Eq. (25)-(27).

It is worth noting that by letting the degrees of freedom ν^* to go to infinity, we can recover the Gaussian distribution with the same parameters. If the data contain outliers, the degrees of freedom parameter ν^* are weak and the mean and covariance of the data are appropriately weighted in order not to take into account the outliers.

Acknowledgements The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. 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.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, NV (2016)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii (2017)
4. Choutas, V., Weinzaepfel, P., Revaud, J., Schmid, C.: PoTion: Pose motion representation for action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT (2018)
5. Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(12), 1553–1566 (2004)
6. Crasto, N., Weinzaepfel, P., Alahari, K., Schmid, C.: Mars: Motion-augmented RGB stream for action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach, CA (2019)
7. De Geest, R., Tuytelaars, T.: Modeling temporal structure with LSTM for online action detection. In: Proc. IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, NV/CA (2018)
8. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Boston, MA (2015)
9. Feichtenhofer, C., Pinz, A., Wildes, R.P., Zisserman, A.: What have we learned from deep representations for action recognition? In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT (2018)

10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, NV (2016)
11. Fouad, S., Tino, P., Raychaudhury, S., Schneider, P.: Incorporating privileged information through metric learning. IEEE Transactions on Neural Networks and Learning Systems **24**(7), 1086–1098 (2013)
12. Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Attribute learning for understanding unstructured social activity. In: Proc. 12$^{\text{th}}$ European Conference on Computer Vision, *Lecture Notes in Computer Science*, vol. 7575. Florence, Italy (2012)
13. Gao, Z., Li, S., Zhu, Y., Wang, C., Zhang, H.: Collaborative sparse representation leaning model for rgbd action recognition. Journal of Visual Communication and Image Representation **48**, 442–452 (2017)
14. Gao, Z., Xuan, H., Zhang, H., Wan, S., Choo, K.R.: Adaptive fusion and category-level dictionary learning model for multiview human action recognition. IEEE Internet of Things Journal **6**(6), 9280–9293 (2019)
15. Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: Proc. European Conference on Computer Vision. Munich, Germany (2018)
16. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R.J., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proc. IEEE International Conference on Computer Vision. Sydney, Australia (2013)
17. Hardoon, D.R., Szedmak, S.R., Shawe-Taylor., J.R.: Canonical correlation analysis: an overview with application to learning methods. Neural Computation **16**(12), 2639–2664 (2004)
18. Hastie, T., Rosset, S., Tibshirani, R., Zhu, J.: The entire regularization path for the support vector machine. Journal of Machine Learning Research **5**, 1391–1415 (2004)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778. Las Vegas, NV (2016)
20. Hoai, M., Zisserman, A.: Talking heads: Detecting humans and recognizing their interactions. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Columbus, OH (2014)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
22. Jin, L., Li, Z., Tang, J.: Deep semantic multimodal hashing network for scalable image-text and video-text retrievals. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2020). DOI 10.1109/TNNLS.2020.2997020
23. Kakadiaris, I., Sarafianos, N., Nikou, C.: Show me your body: Gender classification from still images. In: Proc. IEEE International Conference on Image Processing. Phoenix, AZ (2016)
24. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Columbus, OH (2014)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
26. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proc. British Machine Vision Conference. University of Leeds, Leeds, UK (2008)
27. Komodakis, N., Tziritas, G.: Image completion using efficient belief propagation via priority scheduling and dynamic pruning. IEEE Transactions on Image Processing **16**(11), 2649–2661 (2007)
28. Kotz, S., Nadarajah, S.: Multivariate t distributions and their applications. Cambridge University Press, Cambridge, New York, Madrid (2004)
29. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2-3), 107–123 (2005)
30. Li, C., Zhong, Q., Xie, D., Pu, S.: Collaborative spatiotemporal feature learning for video action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach, CA (2019)
31. Li, Y., Li, Y., Vasconcelos, N.: RESOUND: towards action recognition without representation bias. In: Proc. European Conference on Computer Vision. Munich, Germany (2018)

32. Liu, A., Su, Y., Jia, P., Gao, Z., Hao, T., Yang, Z.: Multiple/single-view human action recognition via part-induced multitask structural learning. IEEE Transactions on Cybernetics **45**(6), 1194–1208 (2015)
33. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO (2011)
34. Lopez-Paz, D., Bottou, L., Schölkopf, B., Vapnik, V.: Unifying distillation and privileged information. In: Proc. 5$^{\text{th}}$ International Conference on Learning Representations. San Juan, Puerto Rico (2016)
35. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal on Computer Vision **60**(2), 91–110 (2004)
36. Luo, Z., Hsieh, J.T., Jiang, L., Carlos Niebles, J., Fei-Fei, L.: Graph distillation for action detection with privileged modalities. In: Proc European Conference on Computer Vision. Munich, Germany (2018)
37. Luvizon, D.C., Picard, D., Tabia, H.: 2D/3D pose estimation and action recognition using multitask deep learning. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT (2018)
38. Marín-Jiménez, M.J., noz Salinas, R.M., Yeguas-Bolivar, E., de la Blanca, N.P.: Human interaction categorization by using audio-visual cues. Machine Vision and Applications **25**(1), 71–84 (2014)
39. Nocedal, J., Wright, S.J.: Numerical optimization, 2nd edn. Springer series in operations research and financial engineering. Springer, New York, NY (2006)
40. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Proc. Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada (2009)
41. Patron-Perez, A., Marszalek, M., Reid, I., Zisserman, A.: Structured learning of human interactions in TV shows. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(12), 2441–2453 (2012)
42. Pechyony, D., Vapnik, V.: On the theory of learnining with privileged information. In: Proc. Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada (2010)
43. Peel, D., Mclachlan, G.J.: Robust mixture modelling using the t distribution. Statistics and Computing **10**, 339–348 (2000)
44. Perrett, T., Damen, D.: DDLSTM: dual-domain LSTM for cross-dataset action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach, CA (2019)
45. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T.: Hidden conditional random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(10), 1848–1852 (2007)
46. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River, NJ (1993)
47. Ramanathan, V., Liang, P., Fei-Fei, L.: Video event understanding using natural language descriptions. In: Proc. IEEE International Conference on Computer Vision. Sydney, Australia (2013)
48. Ramanathan, V., Yao, B., Fei-Fei, L.: Social role discovery in human events. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Portland, OR (2013)
49. Serra-Toro, C., Traver, V.J., Pla, F.: Exploring some practical issues of svm+: Is really privileged information that helps? Pattern Recognition Letters **42**(0), 40–46 (2014)
50. Shao, J., Kang, K., Loy, C.C., Wang, X.: Deeply learned attributes for crowded scene understanding. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Boston, MA (2015)
51. Sharmanska, V., Quadrianto, N., Lampert, C.H.: Learning to rank using privileged information. In: Proc. IEEE International Conference on Computer Vision. Sydney, Australia (2013)
52. Smailis, C., Vrigkas, M., Kakadiaris, I.A.: Recaspia: Recognizing carrying actions in single images using privileged information. In: Proc. 26th IEEE International Conference on Image Processing, pp. 26–30. Taipei, Taiwan (2019)
53. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1–1 (2014)

54. Teo, C.H., Smola, A.J., Vishwanathan, S.V.N., Le, Q.V.: A scalable modular convex solver for regularized risk minimization. In: Proc. ACM International Conference on Knowledge Discovery and Data Mining. San Jose, CA (2007)
55. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proc. IEEE International Conference on Computer Vision, pp. 4489–4497. Santiago, Chile (2015)
56. Tsai, Y.H.H., Divvala, S., Morency, L.P., Salakhutdinov, R., Farhadi, A.: Video relationship reasoning using gated spatio-temporal energy graph. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach, CA (2019)
57. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural Networks **22**(5–6), 544–557 (2009)
58. Vrigkas, M., Kazakos, E., Nikou, C., Kakadiaris, I.A.: Inferring human activities using robust privileged probabilistic learning. In: Proc. IEEE International Conference on Computer Vision Workshops. Venice, Italy (2017)
59. Vrigkas, M., Mastora, E., Nikou, C., Kakadiaris, I.A.: Robust incremental hidden conditional random fields for human action recognition. In: Proc. 13th International Symposium on Visual Computing, pp. 126–136. Las Vegas, NV (2018)
60. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: Classifying behavioral attributes using conditional random fields. In: Proc. 8$^{\text{th}}$ Hellenic Conference on Artificial Intelligence, *Lecture Notes in Computer Science*, vol. 8445. Ioannina, Greece (2014)
61. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and Artificial Intelligence **2**(28), 1–26 (2015). DOI 10.3389/frobt.2015.00028
62. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: Active privileged learning of human activities from weakly labeled samples. In: Proc. 23$^{\text{rd}}$ IEEE International Conference on Image Processing. Phoenix, AZ (2016)
63. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: Exploiting privileged information for facial expression recognition. In: Proc. IEEE International Conference on Biometrics. Halmstad, Sweden (2016)
64. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: Identifying human behaviors using synchronized audio-visual cues. IEEE Transactions on Affective Computing **8**(1), 54–66 (2017)
65. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proc. IEEE International Conference on Computer Vision. Sydney, Australia (2013)
66. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Boston, MA (2015)
67. Wang, S., He, M., Zhu, Y., He, S., Liu, Y., Ji, Q.: Learning with privileged information using Bayesian networks. Frontiers of Computer Science **9**(2), 185–199 (2015)
68. Wang, Y., Mori, G.: Hidden part models for human action recognition: probabilistic versus max margin. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(7), 1310–1323 (2011)
69. Wang, Z., Gao, T., Ji, Q.: Learning with hidden information using a max-margin latent variable model. In: Proc. International Conference on Pattern Recognition. Stockholm, Sweden (2014)
70. Wang, Z., Ji, Q.: Classifier learning with hidden information. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Boston, MA (2015)
71. Yan, A., Wang, Y., Li, Z., Qiao, Y.: PA3D: Pose-action 3D machine for video recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach, CA (2019)
72. Yuan, S., Stenger, B., Kim, T.K.: 3D hand pose estimation from RGB using privileged learning with depth data. In: Proc. IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea (2019)
73. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: 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. Rhode Island (2012)
74. Zhu, W., Hu, J., Sun, G., Cao, X., Qiao, Y.: A key volume mining deep framework for action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Las Vegas, NV (2016)
75. Zhu, Y., Long, Y., Guan, Y., Newsam, S., Shao, L.: Towards universal representation for unseen action recognition. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT (2018)