Inferring Human Activities Using Robust Privileged Probabilistic Learning

Michalis Vrigkas[1] Evangelos Kazakos[2] Christophoros Nikou[2] Ioannis A. Kakadiaris[1]
[1]Computational Biomedicine Lab, University of Houston, Houston, TX, USA
[2]Dept. Computer Science & Engineering, University of Ioannina, Ioannina, Greece

Abstract

Classification models may often suffer from "structure imbalance" between training and testing data that may occur due to the deficient data collection process. This imbalance can be represented by the learning using privileged information (LUPI) paradigm. In this paper, we present a supervised probabilistic classification approach that integrates LUPI into a hidden conditional random field (HCRF) model. The proposed model is called LUPI-HCRF and is able to cope with additional information that is only available during training. Moreover, the proposed method employes Student's t-distribution to provide robustness to outliers by modeling the conditional distribution of the privileged information. Experimental results in three publicly available datasets demonstrate the effectiveness of the proposed approach and improve the state-of-the-art in the LUPI framework for recognizing human activities.

1. Introduction

The rapid development of human activity recognition systems for applications such as surveillance and human-machine interactions [5, 31] brings forth the need for developing new learning techniques. Learning using privileged information (LUPI) [18, 28, 34] has recently generated considerable research interest. The insight of LUPI is that one may have access to additional information about the training samples, which is not available during testing.

Despite the impressive progress that has been made in recognizing human activities, the problem still remains challenging. First, constructing a visual model for learning and analyzing human movements is difficult. The large intra-class variabilities or changes in appearance make the recognition problem difficult to address. Finally, the lack of informative data or the presence of misleading information may lead to ineffective approaches.

We address these issues by presenting a probabilistic approach, which is able to learn human activities by exploiting additional information about the input data, that may reflect on auxiliary properties about classes and members of the classes of the training data (Fig. 1). In this context, we employ a new learning method based on hidden conditional random fields (HCRFs) [24], called LUPI-HCRF, which can efficiently manage dissimilarities in input data, such as noise, or missing data, using a Student's t-distribution. 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 [23].

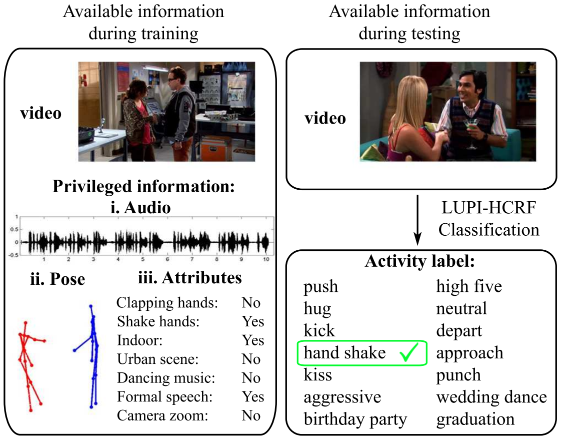


Figure 1. Robust learning using privileged information. Given a set of training examples and a set of additional information about the training samples (left) our system can successfully recognize the class label of the underlying activity without having access to the additional information during testing (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 LUPI-HCRF model.

The main contributions of our work can be summarized in the following points. First, we developed a probabilistic human activity recognition method that exploits privileged information based on HCRFs to deal with missing or incomplete data during testing. Second, contrary to previous methods, which may be sensitive to outlying data measurements, we propose a robust framework by employing a Student's t-distribution to attain robustness against outliers. Finally, we emphasize the generic nature of our approach to cope with samples from different modalities.

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 [27, 36]. Earlier approaches use different kinds of modalities, such as audio information, as additional information to construct better classification models for activity recognition [32].

A shared representation of human poses and visual information has also been explored [40]. Several kinematic constraints for decomposing human poses into separate limbs have been explored to localize the human body [4]. However, identifying which body parts are most significant for recognizing complex human activities still remains a challenging task [16]. Much focus has also been given in recognizing human activities from movies or TV shows by exploiting scene contexts to localize and understand human interactions [10, 22]. The recognition accuracy of such complex videos can also be improved by relating textual descriptions and visual context to a unified framework [26].

Recently, intermediate semantic features representation for recognizing unseen actions during training has been proposed [17, 38]. These features are learned during training and enable parameter sharing between classes by capturing the correlations between frequently occurring low-level features [1]. Instead of learning one classifier per attribute, a two-step classification method has been proposed by Lampert *et al.* [14]. Specific attributes are predicted from pre-trained classifiers and mapped into a class-level score.

Recent methods that exploited deep neural networks have demonstrated remarkable results in large-scale datasets. Donahue *et al.* [6] proposed a recurrent convolutional architecture, where recurrent long-term models are connected to convolutional neural networks (CNNs) that can be jointly trained to simultaneously learn spatio-temporal dynamics. Wang *et al.* [37] proposed a new video representation that employs CNNs to learn multi-scale convolutional feature maps. Tran *et al.* [33] introduced a 3D ConvNet architecture that learns spatio-temporal features using 3D convolutions. 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.* [2]. Feichtenhofer *et al.* [7] introduced a novel architecture for two stream ConvNets and studied different ways for spatio-temporal fusion of the ConvNet towers. Zhu *et al.* [41] argued that videos contain one or more key volumes that are discriminative and most volumes are irrelevant to the recognition process.

The LUPI paradigm was first introduced by Vapnik and Vashist [34] as a new classification setting to model based on a max-margin framework, called SVM+. 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.* [30]. Wand and Ji [39] proposed two different loss functions that exploit privileged information and can be used with any classifier. Recently, a combination of the LUPI framework and active learning has been explored by Vrigkas *et al.* [35] to classify human activities in a semi-supervised scheme.

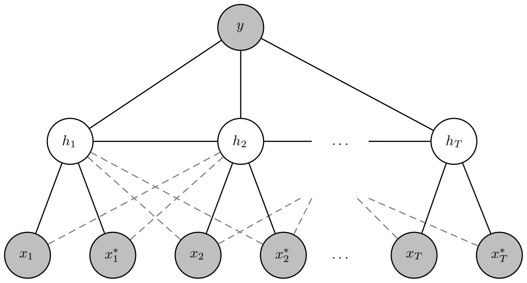


Figure 2. Graphical representation of the chain structure model. The grey nodes are the observed features (x_i), the privileged information (x_i^*), and the unknown labels (y), respectively. The white nodes are the unobserved hidden variables (h).

3. Robust privileged probabilistic learning

Our method uses HCRFs, 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 activity 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) [12].

3.1. LUPI-HCRF model formulation

We consider a labeled dataset 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 $\mathcal{Y}$. Also, the 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 what follows, we omit indices i and j for simplicity.

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

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

where $\mathbf{w} = [\boldsymbol{\theta}, \boldsymbol{\omega}]$ is a vector of model parameters, and $\mathbf{h} = \{h_1, h_2, \dots, h_T\}$, 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 [24]. 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:

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

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:

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

where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ are the unary and the pairwise weights, respectively, that need to be learned. The unary potential does not depend on more than two hidden variables h_j and h_k, and the 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:

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

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:

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

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

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

Finally, the privileged feature function, which models the relationship between the hidden variables h_j is defined by:

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

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:

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

3.2. Parameter learning and inference

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

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

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. The second term is a Gaussian prior with variance σ^2 and works as a regularizer. The loss function is optimized using the limited-memory BFGS (LBFGS) method [20] to minimize the negative log-likelihood of the data.

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 $\mathbf{x}^*$:

$$\begin{aligned} y &= \arg\max_y p(y|\mathbf{x}; \mathbf{w}) \\ &= \arg\max_y \sum_{\mathbf{h}} \sum_{\mathbf{x}^*} p(y, \mathbf{h}|\mathbf{x}, \mathbf{x}^*; \mathbf{w}) p(\mathbf{x}^*|\mathbf{x}; \mathbf{w}) . \end{aligned} \quad (10)$$

To efficiently cope with outlying measurements about the training data, we consider that the training samples $\mathbf{x}$ and $\mathbf{x}^*$ jointly follow a Student's t-distribution. 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, with mean vector μ^*, covariance matrix Σ^* and $\nu^* \in [0, \infty)$ corresponds to the degrees of freedom of the distribution [13]. 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. An approximate inference is employed for estimation of the marginal probability (Eq. (10)) by applying the LBP algorithm [12].

4. Multimodal Feature Fusion

One drawback of combining features of different modalities is the different frame rate that each modality may have. Thus, instead of directly combining multimodal features together one may employ canonical correlation analysis (CCA) [9] to exploit the correlation between the different modalities by projecting them onto a common subspace

Algorithm 1 Robust privileged probabilistic leaning

```
Input: Original data $\mathbf{x}$, privileged data $\mathbf{x}^*$, class labels $y$
Perform feature extraction from both $\mathbf{x}$ and $\mathbf{x}^*$
Project $\mathbf{x}$ and $\mathbf{x}^*$ onto a common space using Eq. (11)
$\mathbf{w}^* \leftarrow \arg\min_{\mathbf{w}} (-L(\mathbf{w}))$ /* Train LUPI-HCRF on $\mathbf{x}$ and $\mathbf{x}^*$ using Eq. (9) */
$\hat{y} \leftarrow \arg\max_{y} p(y|\mathbf{x};\mathbf{w})$ /* Marginalize over $\mathbf{h}$ and $\mathbf{x}^*$ using Eq. (10) */
Output: Predicted labels $\hat{y}$
```

so that the correlation between the input vectors is maximized in the projected space. In this paper, we followed a different approach. Our model is able to 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 [21]. Let $\mathbf{x} \in \mathbb{R}^{M\times d}$ be the input raw data and $\mathbf{a} \in \mathbb{R}^{M\times p}$ be the set of semantic attributes (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:

$$\arg\min_{\boldsymbol{\gamma}} \|\mathbf{x}\boldsymbol{\gamma} - \mathbf{a}\|^2 + \eta\|\boldsymbol{\gamma}\|^2 \,, \tag{11}$$

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 $\mathbf{x}\boldsymbol{\gamma} = \mathbf{a}$, Eq. (11) has a closed form solution $\boldsymbol{\gamma} = \left(\mathbf{x}^T\mathbf{x} + \eta I\right)^{-1}\mathbf{x}^T\mathbf{a}$, where I is the identity matrix. Note that the minimization of Eq. (11) 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.

5. Experiments

Datasets: The TV human interaction (TVHI) [22] dataset consists of 300 videos and contains four kinds of interactions. The SBU Kinect Interaction (SBU) [40] dataset is a collection of approximately 300 videos that contain eight different interaction classes. Finally, the unstructured social activity attribute (USAA) [8] dataset includes eight different semantic class videos of social occasions and contains around 100 videos per class for training and testing.

5.1. Implementation details

Feature selection: For the evaluation of our method, we used spatio-temporal interest points (STIP) [15] as our base video representation. These features were selected because they can capture salient visual motion patterns in an efficient and compact way. In addition, for the TVHI dataset, we also used the provided annotations, which are related to the locations of the persons in each video clip. For this dataset, we used audio features as privileged information. More specifically, we employed the mel-frequency cepstral coefficients (MFCC) [25] features, resulting in a collection of 13 MFCC coefficients, and their first and second order derivatives forming a 39-dimensional feature vector.

Dataset	Features (dimension)	Regular	Privileged
TVHI [22]	STIP (162)	✓	
	Head orientations (2)	✓	
	MFCC (39)		✓
SBU [40]	STIP (162)	✓	
	Pose (15)		✓
USAA [8]	STIP (162)	✓	
	SIFT (128)	✓	
	MFCC (39)	✓	
	Semantic attributes (69)		✓

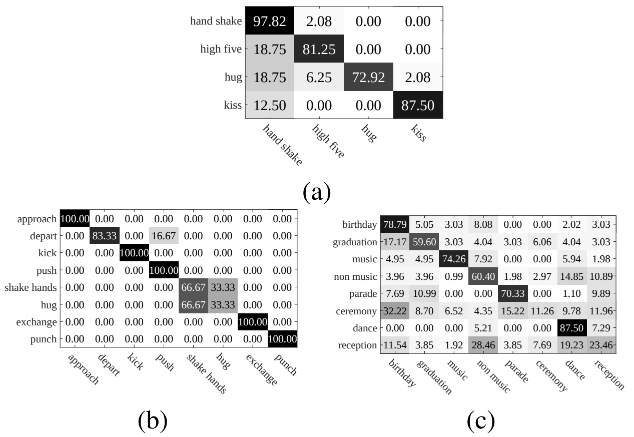
Table 1. Types of features used for human activity recognition for each dataset. The numbers in parentheses indicate the dimension of the features. The checkmark corresponds to the usage of the specific information as regular or privileged. Privileged features are used only during training.

Furthermore, for the SBU dataset, as privileged information, we used the positions of the locations of the joints for each person in each frame, and six more feature types concerning joint distance, joint motion, plane, normal plane, velocity, and normal velocity as described by Yun *et al.* [40]. Finally, for the USAA dataset we used the provided attribute annotation as privileged information to characterize each class with a feature vector of semantic attributes. As a representation of the video data, we used the provided SIFT [19], STIP, and MFCC features. Table 1 summarizes all forms of features used either as regular or privileged for each dataset during training and testing.

Model selection: The model in Fig. 2 was trained by varying the number of hidden states from 4 to 20, with a maximum of 400 iterations for the termination of LBFGS. The L_2 regularization scale term σ was searched within $\{10^{-3}, \ldots, 10^{3}\}$ and 5-fold cross validation was used to split the datasets into training and test sets, and the average results over all the examined configurations are reported.

5.2. Results and discussion

Classification with hand-crafted features: We compare the results of our LUPI-HCRF method with the state-of-the-art SVM+ method [34] and other baselines that incorporate the LUPI paradigm. Also, to demonstrate the efficacy of robust privileged information to the problem of human activity recognition, we compared it with ordinary



(a)

(b)

(c)

Figure 3. Confusion matrices for the classification results of the proposed LUPI-HCRF approach for (a) the TVHI [22], (b) the SBU [40], and (c) the USAA [8] datasets.

SVM and HCRF, as if they could access both the original and the privileged information at test time. This means that we do not differentiate between regular and privileged information, but use both forms of information as regular to infer the underlying class label instead. Also, for the SVM+ and SVM, we consider a one-versus-all decomposition of multi-class classification scheme and average the results for every possible configurations. Finally, the optimal parameters for the SVM and SVM+ were selected using cross validation.

The resulting confusion matrices for all datasets are depicted in Fig. 3. For the TVHI and SBU datasets, the classification errors between different classes are relatively small, as only a few classes are strongly confused with each other. For the USAA dataset, the different classes may be strongly confused (*e.g.*, the class *wedding ceremony* is confused with the class *graduation party*) as the dataset has large intra-class variabilities, while the different classes may share the same attribute representation as different videos may have been captured under similar conditions.

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 in Table 2. The combination of regular and privileged features has considerably increased the recognition accuracy to much higher levels than using each modality separately. If only privileged information is used as regular for classification, the recognition accuracy is notably lower than when using visual information for the classification task. Thus, we may come to the conclusion that finding proper privileged information is not always a straightforward problem.

Table 3 compares the proposed approach with state-of-the-art methods on the human activity classification task on the same datasets. The results indicate that our approach improved the classification accuracy. On TVHI, we significantly managed to increase the classification accuracy by approximately 10% with respect to the SVM+ approach. Also, the improvement of our method with respect to SVM+ was about 6% and 10% for the SBU and USAA datasets, respectively. This indicates the strength of the LUPI paradigm and demonstrates the need for additional information.

Dataset	Regular	Privileged	Accuracy (%)
TVHI [22]	visual	✗	60.9
	audio	✗	35.9
	visual+audio	✗	81.3
	visual	audio	**84.9**
SBU [40]	visual	✗	69.8
	pose	✗	62.5
	visual+pose	✗	81.4
	visual	pose	**85.4**
USAA [8]	visual	✗	55.5
	sem. attributes	✗	37.4
	visual+sem. attributes	✗	54.0
	visual	sem. attributes	**58.1**

Table 2. Comparison of feature combinations for classifying human activities and events on TVHI [22], SBU [40] and USAA [8] datasets. Robust privileged information seems to work in favor of the classification task. The crossmark corresponds to the absence of privileged information during training.

Hand-crafted features			
Method	TVHI	SBU	USAA
Wang and Schmid [36]	76.1 ± 0.4	79.6±0.4	55.5 ± 0.1
Wang and Ji [39]	74.8 ± 0.2	62.4 ± 0.3	48.5 ± 0.2
Sharmanska *et al.* [30]	65.2 ± 0.1	56.3 ± 0.2	56.3 ± 0.2
SVM [3]	75.9 ± 0.6	79.4 ± 0.4	47.4 ± 0.1
HCRF [24]	81.3 ± 0.7	81.4 ± 0.8	54.0 ± 0.8
SVM+ [34]	75.0 ± 0.2	79.4 ± 0.3	48.5 ± 0.1
LUPI-HCRF	$\mathbf{84.9} \pm 0.8$	$\mathbf{85.4} \pm 0.4$	$\mathbf{58.1} \pm 1.4$

Table 3. Comparison of the classification accuracies (%) on TVHI, SBU and USAA datasets using hand-crafted features.

Classification with CNN features: In our experiments, we used CNNs for both end-to-end classification and feature extraction. In both cases, we employed the pre-trained model of Tran *et al.* [33], which is a 3D ConvNet. We selected this model because it was trained on a very large dataset, namely Sports 1M dataset [11], which provides very good features for the activity recognition task, especially in our case where the size of the training data is small.

For the SBU dataset, which is a fairly small dataset, only a few parameters had to be trained to avoid overfitting. We replaced the fully-connected layer of the pre-trained model with a new fully-connected layer of size 1024 and retrained the model by adding a softmax layer on top of it. For the TVHI dataset, we fine-tuned the last group of convolutional layers, while for USAA, 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 optimization process, we used mini-batch gradient descent (SGD) with momentum. The size of the mini-batch was set to 16

	CNN features		
Method	TVHI	SBU	USAA
CNN (end-to-end) [33]	60.5 ± 1.1	94.2 ± 0.8	67.4 ± 0.6
SVM [3]	90.0 ± 0.3	92.8 ± 0.2	91.9 ± 0.3
HCRF [24]	89.6 ± 0.5	91.1 ± 0.4	91.6 ± 0.8
SVM+ [34]	92.5 ± 0.4	94.8 ± 0.3	92.3 ± 0.3
LUPI-HCRF	$\mathbf{93.2} \pm 0.6$	$\mathbf{94.9} \pm 0.7$	$\mathbf{93.9} \pm 0.9$

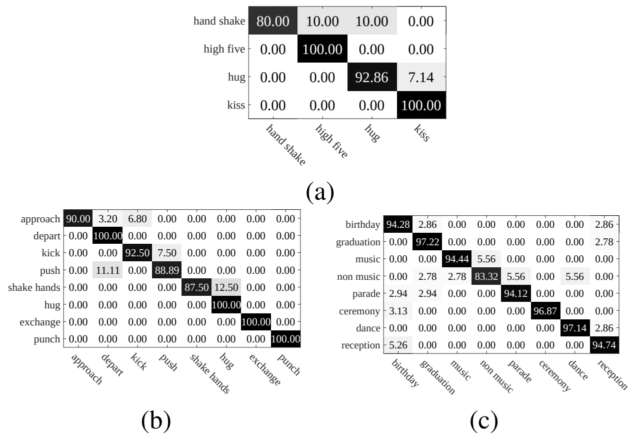
Table 4. Comparison of the classification accuracies (%) on TVHI, SBU and USAA datasets using CNN features.

with a constant momentum of 0.9. For the SBU dataset, 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 1000. 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 probability 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 classification, 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 CNN features from the newly added fully-connected layer.

Table 4 summarizes the comparison of LUPI-HCRF with state-of-the-art methods using the features extracted from the CNN model, and end-to-end learning of the CNN model using softmax loss for the classification. The improvement of accuracy, compared to the classification based on the traditional features (Table 3), indicates that CNNs may efficiently extract informative features without any need to hand design them. We may observe that privileged information works in favor of the classification task in all cases. LUPI-HCRF achieves notably higher recognition accuracy with respect to the HCRF model and the SVM+ approaches. Moreover, for both TVHI and USAA datasets, when LUPI-HCRF is compared to the end-to-end CNN model, it achieved an improvement of approximately 33% and 27%, respectively. This huge improvement can be explained by the fact that the CNN model uses a very simple classifier in the softmax layer, while LUPI-HCRF is a more sophisticated model that can efficiently handle sequential data in a more principled way.

The corresponding confusion matrices using the CNN-based features, are depicted in Fig. 4. The combination of privileged information with the information learned from the CNN model feature representation resulted in very small inter- and intra-class classification errors for all datasets.

Discussion: In general, our method is able to robustly use privileged information in a more efficient way than its SVM+ counterpart by exploiting the hidden dynamics between the videos and the privileged information. However, selecting which features can act as privileged information is not straightforward. The performance of LUPI-based classifiers relies on the delicate relationship between the regular and the privileged information. Tables 3 and 4 show that SVM and HCRF perform worse than LUPI-HCRF. This is because at training time privileged information comes as ground truth but at test time it is not. Also, privileged information is costly or difficult to obtain with respect to producing additional regular training examples [29].



	hand shake	high five	hug	kiss
hand shake	80.00	10.00	10.00	0.00
high five	0.00	100.00	0.00	0.00
hug	0.00	0.00	92.86	7.14
kiss	0.00	0.00	0.00	100.00

(a)

	approach	depart	kick	push	shake hands	hug	exchange	punch
approach	90.00	3.20	6.80	0.00	0.00	0.00	0.00	0.00
depart	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	92.50	7.50	0.00	0.00	0.00	0.00
push	0.00	11.11	0.00	88.89	0.00	0.00	0.00	0.00
shake hands	0.00	0.00	0.00	0.00	87.50	12.50	0.00	0.00
hug	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
exchange	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
punch	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

(b)

	birthday	graduation	music	non music	parade	ceremony	dance	reception
birthday	94.28	2.86	0.00	0.00	0.00	0.00	0.00	2.86
graduation	0.00	97.22	0.00	0.00	0.00	0.00	0.00	2.78
music	0.00	0.00	94.44	5.56	0.00	0.00	0.00	0.00
non music	0.00	2.78	2.78	83.32	5.56	0.00	5.56	0.00
parade	2.94	2.94	0.00	0.00	94.12	0.00	0.00	0.00
ceremony	3.13	0.00	0.00	0.00	0.00	96.87	0.00	0.00
dance	0.00	0.00	0.00	0.00	0.00	0.00	97.14	2.86
reception	5.26	0.00	0.00	0.00	0.00	0.00	0.00	94.74

(c)

Figure 4. Confusion matrices for the classification results of the proposed LUPI-HCRF approach for (a) the TVHI [22], (b) the SBU [40], and (c) the USAA [8] datasets using the CNN features.

6. Conclusion

In this paper, we addressed the problem of human activity recognition and proposed a novel probabilistic classification model based on robust learning by incorporating a Student's t-distribution into the LUPI paradigm, called LUPI-HCRF. We evaluated the performance of our method on three publicly available datasets and tested various forms of data that can be used as privileged. The experimental results indicated that robust privileged information ameliorates the recognition performance. We demonstrated improved results with respect to the state-of-the-art approaches that may or may not incorporate privileged information.

Acknowledgments

This work was funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund and by the European Commission (H2020-MSCA-IF-2014), under grant agreement No. 656094. 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.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 819–826, Portland, OR, June 2013.

[2] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3034–3042, Las Vegas, NV, June 2016.

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

[4] A. Cherian, J. Mairal, K. Alahari, and C. Schmid. Mixing body-part sequences for human pose estimation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2361–2368, Columbus, OH, June 2014.

[5] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1566, December 2004.

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, Boston, MA, June 2015.

[7] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, Las Vegas, NV, June 2016.

[8] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *Proc.* 12^{th} *European Conference on Computer Vision*, volume 7575 of *Lecture Notes in Computer Science*, pages 530–543, Firenze, Italy, October 2012.

[9] 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):2639–2664, Dec. 2004.

[10] M. Hoai and A. Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 875–882, Columbus, OH, June 2014.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, Columbus, OH, June 2014.

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

[13] S. Kotz and S. Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, Cambridge, 2004.

[14] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 951–958, Miami Beach, Florida, June 2009.

[15] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, September 2005.

[16] I. Lillo, A. Soto, and J. C. Niebles. Discriminative hierarchical modeling of spatio-temporally composable human activities. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 812–819, Columbus, OH, June 2014.

[17] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, Colorado Springs, CO, June 2011.

[18] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. In *Proc. International Conference on Learning Representations*, San Jose, Puerto Rico, May 2016.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 60(2):91–110, Nov. 2004.

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

[21] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *Proc. Advances in Neural Information Processing Systems*, pages 1410–1418, Vancouver, British Columbia, Canada, December 2009.

[22] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in TV shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(12):2441–2453, Dec. 2012.

[23] D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

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

[25] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ, 1993.

[26] V. Ramanathan, P. Liang, and L. Fei-Fei. Video event understanding using natural language descriptions. In *Proc. IEEE International Conference on Computer Vision*, pages 905–912, Sydney, Australia, December 2013.

[27] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Portland, OR, June 2013.

[28] N. Sarafianos, C. Nikou, and I. A. Kakadiaris. Predicting privileged information for height estimation. In *Proc. Inter-*

national Conference on Pattern Recognition, Cancun, Mexico, December 2016.

[29] 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.

[30] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Learning to rank using privileged information. In *Proc. IEEE International Conference on Computer Vision*, pages 825–832, Sydney, Australia, December 2013.

[31] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: an experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1442–1468, 2014.

[32] Y. Song, L. P. Morency, and R. Davis. Multi-view latent variable discriminative models for action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, June 2012.

[33] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proc. IEEE International Conference on Computer Vision*, pages 4489–4497, Santiago, Chile, December 2015.

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

[35] M. Vrigkas, C. Nikou, and I. A. Kakadiaris. Active privileged learning of human activities from weakly labeled samples. In *Proc. 23rd IEEE International Conference on Image Processing*, pages 3036–3040, Phoenix, AZ, September 2016.

[36] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. IEEE International Conference on Computer Vision*, pages 3551–3558, Sydney, Australia, December 2013.

[37] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, Boston, MA, June 2015.

[38] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proc.* 11*th European Conference on Computer Vision: Part V*, pages 155–168, Heraklion, Crete, Greece, 2010.

[39] Z. Wang and Q. Ji. Classifier learning with hidden information. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4969–4977, Boston, MA, June 2015.

[40] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, Providence, Rhode Island, June 2012.

[41] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao. A key volume mining deep framework for action recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1991–1999, Las Vegas, NV, June 2016.