ACTIVE PRIVILEGED LEARNING OF HUMAN ACTIVITIES FROM WEAKLY LABELED SAMPLES

Michalis Vrigkas[1] *Christophoros Nikou*[1,2] *Ioannis A. Kakadiaris*[2]

[1] Dept. Computer Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece
{mvrigkas, cnikou}@cs.uoi.gr
[2] Dept. Computer Science, University of Houston, 4800 Calhoun Rd, Houston, TX 77204, USA
ioannisk@uh.edu

ABSTRACT

In many human activity recognition systems the size of the unlabeled training data may be significantly large due to expensive human effort required for data annotation. Moreover, the insufficient data collection process from heterogenous sources may cause dissimilarities between training and testing data. To address these limitations, a novel probabilistic approach that combines learning using privileged information (LUPI) and active learning is proposed. A pool-based privileged active learning approach is presented for semi-supervising learning of human activities from multimodal labeled and unlabeled data. Both uncertainty and distance from the decision boundary are used as a query inference strategies for selecting an unlabeled observation and query its label. Experimental results in four publicly available datasets demonstrate that the proposed method can identify with high accuracy complex human activities.

Index Terms— Learning using privileged information, active learning, hidden conditional random fields, activity recognition

1. INTRODUCTION

Standard human activity classification systems assume that both training and testing sequences represent similar types of information [1]. However, in real-world applications, this may not always be possible due to data acquisition constraints. To overcome this limitation, Vapnik and Vashist [2] introduced the learning using privileged information (LUPI) paradigm. Their method is based on a max-margin classification scheme, called SVM+, and encodes additional information about the training data, which is accessible only during training but never during testing. The goal of privileged information is to build a stronger classifier, that is able to cope with incomplete data during testing. The applications of the LUPI paradigm may vary from clustering [3] and textual description [4], to facial expression [5] and human activity [6] recognition.

In the literature, many variants of SVM+ have been proposed, including SVM+ with $L1$ regularization [7], multi-task SVM+ [8] and risk bound minimization [9]. Although SVM+ and its variants promise good classification results with respect to the standard SVM, they require tuning more parameters for optimizing the loss function for the regular and the privileged feature space.

Niu *et al.* [6] combined multiple instance learning and privileged information to classify human activities and events from web data, while domain adaptation is also considered to address the problem of multimodal data association. Wang *et al.* [10] exploited privileged information using a latent max-margin model. Hidden variables were used as an additional level of abstraction to propagate privileged information and learn the corresponding label.

Most of the recognition systems including LUPI-based classification systems assume that labeled training data are easy to obtain. However, knowing a priori the label of all training examples may not always be feasible for large databases as the cost for manually labeling all samples may be prohibitively large. To address this limitation, active learning has been proposed [11]. The idea of active learning is closely related to semi-supervised learning as during training, labeled and unlabeled data co-exist. The aim of active learning is to actively select the most informative unlabeled samples according to a specified criterion, query their label and use them as training data to construct a stronger classifier. Active learning has been used with several classification models such as SVM [12], conditional random fields [13] and radial basis function networks [14].

An interesting application of active learning is the automatic annotation of ongoing activities in unsegmented video sequences for detecting and localizing human actions [15]. Hasan and Roy-Chowdhury [16] proposed an incremental algorithm for actively learning new actions from streaming videos. However, one of the main problems of active learning is how to define an effective criterion for selecting unlabeled samples [17]. To this end, the same authors [18] combined entropy and mutual information to handle inter and intra-relationships between training data through incremental update of the classification model to learn human activities. Finally, Long *et al.* [19] considered an action recognition method that exploits active learning to cope with multiple and noisy labels.

Previous methods can either handle information that is not available during testing, or cope with missing labels during training but cannot address both problems simultaneously. In this work, we propose a novel classification method that combines the LUPI paradigm and active learning for identifying human activities in a semi-supervised framework using hidden conditional random fields (HCRFs) [20], called active-HCRF+ (a-HCRF+). The proposed method exploits privileged information as an additional input during training to learn the conditional probability distribution between human activities and observations. To reduce tedious human effort in data annotation, an incremental pool-based active learning technique is adopted to actively select unlabeled training samples for which the uncertainty about their actual class label is reduced.

2. ACTIVE PRIVILEGED LEANING

We consider a labeled dataset with N video sequences, which consists of triplets $\mathcal{D} = \{(\mathbf{x}_{i,j}, \mathbf{x}^*_{i,j}, y_i)\}_{i=1}^N$, where $\mathbf{x}_{i,j} \in \mathbb{R}^{M_\mathbf{x} \times T}$ is an observation sequence of length T with $j = 1 \dots T$, which belongs in feature space $\mathcal{X}$. Furthermore, y_i corresponds to a class la-

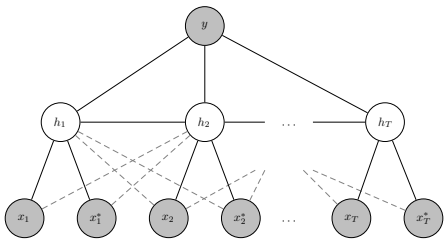

Fig. 1. Graphical representation of the chain structure model. The grey nodes are the observed features (x_i and x_i^*), and the unknown labels (y). The white nodes are the hidden variables (h).

bel defined in a finite label set $\mathcal{Y}$. Also additional information about the observations $\mathbf{x}_i$ is encoded in a feature vector $\mathbf{x}_{i,j}^* \in \mathbb{R}^{M_{\mathbf{x}^*} \times T}$ and belongs to feature space $\mathcal{X}^*$. This information is provided only at the training step and it is not available during testing, while not any assumption about the form of the privileged data is made. In what follows, we omit indices i and j for simplicity.

2.1. a-HCRF+ model formulation

The a-HCRF+ model is defined by a chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (Fig. 1). The proposed model is a member of the exponential family and the probability of the class label given an observation sequence is given by:

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

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

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

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

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

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

The unary potential is expressed by:

$$\begin{aligned} \Phi(y, h_j, \mathbf{x}_j, \mathbf{x}_j^*; \boldsymbol{\theta}) = &\sum_j \phi_1(y, h_j; \boldsymbol{\theta}_1) + \sum_j \phi_2(h_j, \mathbf{x}_j; \boldsymbol{\theta}_2) \\ &+ \sum_j \phi_3(h_j, \mathbf{x}_j^*; \boldsymbol{\theta}_3) , \end{aligned} \quad (4)$$

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

$$\phi_1(y, h_j; \boldsymbol{\theta}_1) = \sum_{\substack{\lambda \in \mathcal{Y} \\ a \in \mathcal{H}}} \sum_{\ell} \boldsymbol{\theta}_{1,\ell} \mathbb{1}(y = \lambda) \mathbb{1}(h_j = a) , \quad (5)$$

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 $\mathbf{x}_j$, is defined by:

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

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

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

The pairwise potential is expressed by:

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

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

2.2. Learning and inference

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

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

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 L_2 regularization Gaussian prior with variance σ^2. The use of hidden variables makes the optimization of Eq. (9) non-convex, thus, a global solution is not guaranteed and we can estimate $\mathbf{w}^*$ that are locally optimal. The loss function is optimized using the limited-memory BFGS (LBFGS) method [21] to minimize the negative log-likelihood of the data.

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

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

In the general case, the training samples $\mathbf{x}$ and $\mathbf{x}^*$ may be considered to be jointly Gaussian, thus the conditional distribution $p(\mathbf{x}^*|\mathbf{x}; \mathbf{w})$ is also a Gaussian distribution. We quantized the continuous space of features to a large number of discrete values to approximate the true value of the marginalization of Eq. (10). However, an exact solution to Eq. (10) is generally intractable. Therefore, approximate inference is employed for estimation of the marginal probability by applying the loopy belief propagation algorithm [22].

2.3. Active learning

In active learning, we suppose that during training we have access to a labeled dataset $\mathcal{L} = \{(\mathbf{x}_{\ell_i}, \mathbf{x}_{\ell_i}^*, y_i)\}_{i=1}^{N_\ell}$, with N_ℓ video sequences and an unlabeled dataset $\mathcal{U} = \{(\mathbf{x}_{u_i}, \mathbf{x}_{u_i}^*)\}_{i=1}^{N_u}$, with N_u

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

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

$$\hat{u} = \arg\max_{u \in \mathcal{U}} \left(-\sum_{y_u} p(y_u|u; \mathbf{w}) \log p(y_u|u; \mathbf{w}) \right) . \quad (11)$$

The second criterion corresponds to the ratio of class posteriors [14]. We estimate the class posterior for each unlabeled observation u and every class. Then, for these two classes that exhibit the largest posterior values $y_1 = \arg\max_{y_u} p(y_u|u; \mathbf{w})$ and $y_2 = \arg\max_{y_u \neq y_1} p(y_u|u; \mathbf{w})$, respectively, we select the unlabeled sample u that minimizes the ratio between the largest class posteriors:

$$\hat{u} = \arg\min_{u \in \mathcal{U}} \frac{p(y_1|u; \mathbf{w})}{p(y_2|u; \mathbf{w})} . \quad (12)$$

The ratio of class posteriors criterion allows to select an observation that lies closer to decision boundary of the learned classifier. Specifically, the main steps of proposed pool-based active learning methodology are summarized in Algorithm 1.

Algorithm 1 Pool-based active learning using a-HCRF+

procedure ACTIVEHCRFPLUS($\mathcal{L}, \mathcal{U}, \mathcal{X}, \mathcal{X}^*, \mathcal{Y}$)
  $\mathbf{w} \leftarrow \arg\min_{\mathbf{w}} (-L(\mathbf{w}))$ ▷ Train a-HCRF+ on $\mathcal{L}$.
  while $\mathcal{U} \neq \emptyset$ **do**
    Select an unlabeled observation $\hat{u}$ according to Eqs. (11) or (12) and query its label y.
    $\mathcal{L} \leftarrow \mathcal{L} \cup \{(\hat{u}, y_u)\}$; ▷ Update labeled dataset $\mathcal{L}$.
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\hat{u}\}$; ▷ Update unlabeled dataset $\mathcal{U}$.
  end while
  $\mathbf{w} \leftarrow \arg\min_{\mathbf{w}} (-L(\mathbf{w}))$ ▷ Update a-HCRF+ parameters.
end procedure

3. EXPERIMENTAL RESULTS

We used four publicly available human activity recognition benchmark datasets. The Parliament dataset [23] contains 228 video sequences of political speeches, belonging in three behavioral categories: friendly, aggressive, and neutral. The TV human interaction (TVHI) dataset [24], is a group of 300 video sequences and contains four kinds of interactions: hand shakes, high fives, hugs, and kisses. The two-person interaction (TPI) dataset [25] consists of approximately 300 video sequences captured by a Microsoft Kinect sensor. The sequences are categorized in eight different interaction classes including approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. Finally, the unstructured social activity attribute (USAA) dataset [26] contains around 100 videos per class for training and testing, while it includes eight different semantic class videos of social occasions such as birthday party, graduation party, music performance, non-music performance, parade, wedding ceremony, wedding dance, and wedding reception.

As video representation for all datasets, we used spatio-temporal interest points (STIP) [27]. Furthermore, for the Parliament and TVHI datasets, we extracted the mel-frequency cepstral coefficients (MFCC) [28] features along with their first and second order derivatives. Audio features are also used as privileged information for these datasets. For the TPI dataset, we used the provided poses as privileged information, and for the USAA dataset we used the provided attribute annotation as privileged information. The number of hidden states was estimated based on cross validation, varying their from 3 to 20. The L_2 regularization scale term σ for was set to 10^k, with $k \in \{-3, \ldots, 3\}$. The proposed model was trained with a maximum of 400 iterations for the termination of the LBFGS optimization method. For each dataset we used 5-fold cross validation to split into training and test sets. Finally, the initial training set was split into labeled and unlabeled set so that the size of the unlabeled set may vary from 10% to 50% of the total size of the original training set and the remaining videos form the labeled training set.

According to which selection criterion is employed (entropy or ratio of class posteriors), we proposed two variants of the method, called a-HCRF+ (entropy) and a-HCRF+ (ratioCP). We compared the proposed method with several baseline methods that may or may not use privileged information and/or active learning. First, we compared it with ordinary SVM [29] and HCRF [20], as if they could access both the original and the privileged information at test time. We also compared with state-of-the-art methods that employ privileged information such as SVM+ [2], the rank transfer SVM+ (rt-SVM+) [4], which exploits a max-margin technique to transfer knowledge from the privileged to the original feature space, and the method of Wang and Ji [5], which exploits a loss inequality regularization (LIR) to address the sensitiveness of the loss function against the inequality constraints. However, these methods do not employ active learning, thus, we also compare with the method of Druck *et al.* [13], which applies generalized expectation criteria such as entropy (GEE) to select the most uncertain samples. Finally, we transformed standard SVM to an active learning based method (a-SVM) using entropy as selection criterion. For the SVM-based methods we consider a one-versus-one decomposition of multi-class classification scheme and average the results for every possible configuration, while the optimal parameters were selected using cross validation.

We assess the impact of privileged active learning by measuring the classification accuracy of both variants of the proposed method with varying number of unlabeled data. The obtained results are depicted in Figure 2. We may observe that for all datasets both pool-based active learning variants (entropy and ratio of class posteriors) always have superior performance than GEE and a-SVM methods as the size of unlabeled training observations increases. Specifically, for the TVHI dataset GEE may perform better only for the a-HCRF+ (ratioCP) variant, while for the USAA dataset a-HCRF+ (ratioCP) and a-SVM achieve similar results. This indicates the strength of the proposed privileged active learning method to recognize human actions from weakly labeled data without loosing accuracy due to the uncertainty of the model about class of each observation.

Detailed results of the proposed method compared with state-of-the-art methods are presented in Table 1. We may observe that for all four datasets the proposed a-HCRF+ (entropy) method outperforms the state-of-the-art. For this variant, the classification performance significantly increased with respect to the LUPI-based SVM+ method for all datasets (e.g., 20% improvement of the Parlia-

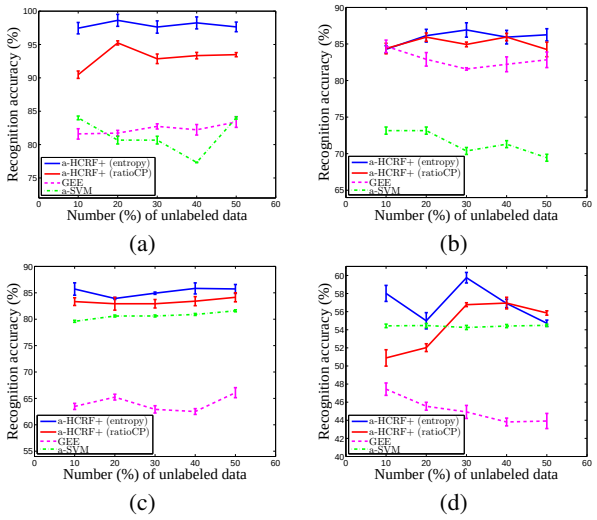

Fig. 2. Comparison of classification accuracies with respect to the number of unlabeled data for (a) the Parliament [23], (b) the TVHI [24], (c) the TPI [25], and (d) the USAA [26] datasets.

Table 1. Comparison of the classification accuracies (%) for the Parliament [23], TVHI [24], TPI [25], and USAA [26] datasets. The results were averaged for all different configurations (mean ± standard deviation).

Method	Parliament [23]	TVHI [24]	TPI [25]	USAA [26]
Methods without privileged information and without active learning				
HCRF [20]	97.6 ± 0.6	81.3 ± 0.7	81.4 ± 0.8	54.0 ± 0.8
SVM [29]	72.6 ± 0.4	75.9 ± 0.6	79.4 ± 0.4	47.4 ± 0.1
Methods without privileged information and with active learning				
GEE [13]	82.3 ± 0.6	83.8 ± 0.8	66.1 ± 0.7	45.4 ± 0.6
a-SVM	80.5 ± 0.3	71.5 ± 0.5	80.6 ± 0.2	54.4 ± 0.2
Methods with privileged information and without active learning				
SVM+ [2]	78.4 ± 0.2	75.0 ± 0.2	79.4 ± 0.3	48.5 ± 0.1
rt-SVM+ [4]	57.7 ± 0.4	65.2 ± 0.1	56.3 ± 0.2	56.3 ± 0.2
LIR [5]	59.2 ± 0.2	74.8 ± 0.2	62.4 ± 0.3	48.5 ± 0.2
Methods with privileged information and with active learning				
a-HCRF+ (entropy)	$\mathbf{98.1} \pm 0.9$	$\mathbf{85.8} \pm 0.5$	$\mathbf{85.2} \pm 0.6$	$\mathbf{56.9} \pm 0.4$
a-HCRF+ (ratioCP)	93.0 ± 0.2	85.1 ± 0.8	83.8 ± 1.0	55.2 ± 0.5

ment dataset). Moreover, significant improvement is obtained, when the proposed method is compared to the active learning counterpart methods. Furthermore, the performance of the a-HCRF+ (ratioCP) variant achieves similar results to its counterpart that uses entropy as a selection criterion. Although the ratio of class posteriors for the Parliament and TVHI datasets may perform worse than standard HCRF model the overall performance is still better than the other methods. This is because of the presence of closely related classes as for some observation close to the decision boundary between two classes the logarithmic ratio of class posteriors may approach zero.

The corresponding confusion matrices for the a-HCRF+ (entropy) variant for the best split for each dataset are shown in Figure 3. It is worth mentioning that for the Parliament and TVHI datasets the classification errors between different classes are relatively small. For the TPI dataset, only a few classes are highly correlated to each other (e.g., the class *shake hands* is confused with the classes *push* and *hug*). On the other hand, the USAA dataset, shows high confusion between the different classes (e.g., *wedding ceremony* is confused with the class *birthday party*). This is because of the large intra-class variabilities, since different classes may have similar attribute representation of human actions.

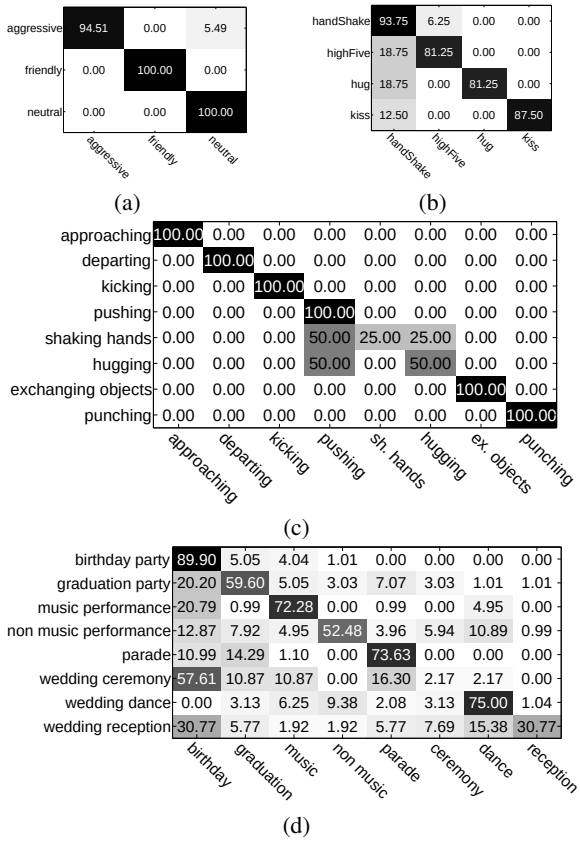

Fig. 3. Confusion matrices for the classification results for the best split of the proposed a-HCRF+ (entropy) variant for (a) the Parliament [23], (b) the TVHI [24], (c) the TPI [25], and (d) the USAA [26] datasets.

4. CONCLUSION

In this paper, the problem of human activity recognition in a semi-supervised framework is investigated. A combination of learning using privileged information and active learning into a unified framework indicated that human actions can effectively be recognized. Moreover, two variants of the proposed a-HCRf+ method were proposed. The fist uses entropy as a measure of uncertainty of the actual class of unlabeled observations and the second selects an unlabeled observation that lies closer to the decision boundary. Several types of auxiliary information were used indicating that the proposed method is not limited to a specific form of privileged information. The experimental results on four different publicly available datasets were very promising and supported the fact that both LUPI and active learning schemes, when used together, achieve superior performance than the state-of-the-art. In future work, we plan to investigate other query selection criteria and how active learning can be used to recognize actions from unsegmented sequences.

5. REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[2] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5–6, pp. 544–557, 2009.

[3] J. Feyereisl and U. Aickelin, "Privileged information for data clustering," *Information Sciences*, vol. 194, no. 0, pp. 4–23, 2012.

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

[5] Z. Wang and Q. Ji, "Classifier learning with hidden information," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 4969–4977.

[6] L. Niu, W. Li, and D. Xu, "Exploiting privileged information from web data for action and event recognition," *International Journal of Computer Vision*, pp. 1–21, 2015.

[7] L. Niu, Y. Shi, and J. Wu, "Learning using privileged information with L-1 support vector machine," in *Proc. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, China, December 2012, vol. 3, pp. 10–14.

[8] Y. Ji and S. Sun, "Multitask multiclass support vector machines: Model and experiments," *Pattern Recognition*, vol. 46, no. 3, pp. 914–924, 2013.

[9] D. Pechyony and V. Vapnik, "On the theory of learnining with privileged information," in *Proc. Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, December 2010, pp. 1894–1902.

[10] Z. Wang, T. Gao, and Q. Ji, "Learning with hidden information using a max-margin latent variable model," in *Proc. International Conference on Pattern Recognition*, Stockholm, Sweden, August 2014, pp. 1389–1394.

[11] B. Settles, "Active learning literature survey," Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[12] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, March 2002.

[13] G. Druck, B. Settles, and A. McCallum, "Active learning by labeling features," in *Proc. Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 81–90.

[14] C. Constantinopoulos and A. Likas, "Semi-supervised and active learning with the probabilistic RBF classifier," *Neurocomputing*, vol. 71, no. 13–15, pp. 2489–2498, 2008.

[15] S. Bandla and K. Grauman, "Active learning of an action detector from untrimmed videos," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 1833–1840.

[16] M. Hasan and A. K. Roy-Chowdhury, "Incremental activity modeling and recognition in streaming videos," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June 2014, pp. 796–803.

[17] S. J. Huang, S. Chen, and Z. H. Zhou, "Multi-label active learning: Query type matters," in *Proc. International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, July 2015, pp. 946–952.

[18] M. Hasan and A. K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *Proc. IEEE International Conference on Computer Vision*, Santiago, Chile, December 2015, pp. 4543–4551.

[19] C. Long, G. Hua, and A. Kapoor, "Active visual recognition with expertise estimation in crowdsourcing," in *Proc. IEEE International Conference on Computer Vision*, Sydney, Australia, December 2013, pp. 3000–3007.

[20] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1848–1852, 2007.

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

[22] N. Komodakis and G. Tziritas, "Image completion using efficient belief propagation via priority scheduling and dynamic pruning," *IEEE Transactions on Image Processing*, vol. 16, no. 11, pp. 2649–2661, November 2007.

[23] Michalis Vrigkas, Christophoros Nikou, and Ioannis A. Kakadiaris, "Classifying behavioral attributes using conditional random fields," in *Proc. 8th Hellenic Conference on Artificial Intelligence*, Ioannina, Greece, May 2014, vol. 8445 of *Lecture Notes in Computer Science*, pp. 95–104.

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

[25] 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*, Rhode Island, USA, June 2012, pp. 28–35.

[26] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Attribute learning for understanding unstructured social activity," in *Proc. 12th European Conference on Computer Vision*, Florence, Italy, October 2012, vol. 7575 of *Lecture Notes in Computer Science*, pp. 530–543.

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

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

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