Robust Incremental Hidden Conditional Random Fields for Human Action Recognition

Michalis Vrigkas[1], Ermioni Mastora[2], Christophoros Nikou[2], and Ioannis A. Kakadiaris[1]

[1] Computational Biomedicine Lab, Department of Computer Science, University of Houston, Houston, TX, USA
{mvrigkas,ikakadia}@central.uh.edu
[2] Department of Computer Science and Engineering, University of Ioannina, Ioannina, Greece
{emastora,cnikou}@cse.uoi.gr

Abstract. Hidden conditional random fields (HCRFs) are a powerful supervised classification system, which is able to capture the intrinsic motion patterns of a human action. However, finding the optimal number of hidden states remains a severe limitation for this model. This paper addresses this limitation by proposing a new model, called robust incremental hidden conditional random field (RI-HCRF). A hidden Markov model (HMM) is created for each observation paired with an action label and its parameters are defined by the potentials of the original HCRF graph. Starting from an initial number of hidden states and increasing their number incrementally, the Viterbi path is computed for each HMM. The method seeks for a sequence of hidden states, where each variable participates in a maximum number of optimal paths. Thereby, variables with low participation in optimal paths are rejected. In addition, a robust mixture of Student's t-distributions is imposed as a regularizer to the parameters of the model. The experimental results on human action recognition show that RI-HCRF successfully estimates the number of hidden states and outperforms all state-of-the-art models.

Keywords: Student's t-distribution · Hidden conditional random fields · Hidden Markov model · Action recognition.

1 Introduction

In recent years, a tremendous amount of human action video recordings has been made available. As a consequence, human action recognition has become a very popular task in computer vision with a wide range of applications such as visual surveillance systems, human-robot interaction, video retrieval, and sports video analysis [21]. The recognition of human actions in videos is a challenging task due to anthropometric differences (size, gender, shape) among subjects and variations in the way, spread, and speed of the action.

Hidden conditional random fields (HCRF) [13] are a generalization of conditional random fields (CRF) [9] and appear to be a very promising approach in

many application domains, due to their ability of relaxing strong independence assumptions and exploiting spatio-temporal variations, via a graphical structure. A hybrid model that consists of a combination of generative and discriminative models to improve the performance of the classical models has been proposed in the literature [1]. Motivated by this approach, Soullard *et al.* [17] introduced an HMM-based weighting in the conditional probability of the HCRF, which constrains the discriminative learning, yielding improved accuracy. On the other hand, Zhang *et al.* [25] used HMMs to make hidden variables "observable" to HCRF so the objective function can be convex. Multi-modal action recognition (i.e., combination of audio and visual information) using HCRFs has also been given great focus [16, 19, 22].

However, previous works define the number of hidden variables in a intuitive manner or with exhaustive search, which is a computationally expensive and time-consuming task. Bousmalis *et al.* [2] introduced infinite hidden conditional random fields (iHCRF), a nonparametric model that estimates the number of hidden variables to recognize human behaviors. The model assumes that the HCRF potentials are sampled directly from a set of hierarchical Dirichlet processes and its hyper-parameters are learned using the sampling that removes hidden variables that are not presented in the samples. Moreover, Bousmalis *et al.* [3] proposed an extension of the previous model, called variational HCRF, which is a generalized framework for infinite HCRF modeling and variational inference in order to converge faster and reduce the computational cost.

Recently, deep leaning methods have shown outstanding results [5]. Although important progress has been made in the fields such as object detection and image classification, the understanding of human actions is still a difficult task due to pose variability, short duration of actions, and ambiguity in human annotations. Sigurdsson *et al.* [15] tried to answer a set of fundamental questions regarding the problem of action recognition to address the aforementioned limitations. Convolutional neural networks (CNNs or ConvNets) [7] are the most widely used approach to simultaneously learn spatio-temporal dynamics of human actions. A novel architecture that that uses spatio-temporal fusion by combining the ConvNet towers was introduced by Feichtenhofer *et al.* [8]. Finally, an on-line frame-pooling method, which extracts only the most important frames that best describe human actions, was proposed by Wang *et al.* [23].

In this work, a robust incremental hidden conditional random field (RI-HCRF) is proposed, which addresses two major issues in standard HCRFs. At first, the proposed model incrementally estimates the optimal number of hidden variables using a Viterbi-based procedure. Additionally, it uses a mixture of Student's t-distributions as prior to the parameters of the model that leads to a model robust to outliers.

2 Model Formulation

The proposed RI-HCRF model is defined by a linear chained structured undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each variable, observed (video frame) and unobserved

(hidden variable), is a node $\mathcal{V}$ in the graphical model $\mathcal{G}$ and any dependencies between them are presented by an edge $\mathcal{E}$. We consider a dataset $\mathcal{D} = \{\mathbf{x}^k, y^k\}_{k=1}^N$ of N labeled observations, where an observation $\mathbf{x}^k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ corresponds to the k^{th} video sequence that consists of T frames, and y^k is the k^{th} class label defined in a finite label set $\mathcal{Y}$. Each observation $\mathbf{x}_i^k$ can be represented by a feature vector $\phi(\mathbf{x}_i^k) \in \mathbb{R}^d$, which is a collection of several features extracted from the i^{th} frame in the video sequence.

The goal of the proposed model is, for a given observation $\mathbf{x}$ and the model's parameter vector $\boldsymbol{\theta}$, to find the most probable label y by maximizing the conditional probability $P(y|\mathbf{x};\boldsymbol{\theta})$. The RI-HCRF model is a member of the exponential family and the conditional probability of the class label given an observation sequence is defined as:

$$P(y|\mathbf{x};\boldsymbol{\theta}) = \sum_h P(y, \mathbf{h}|\mathbf{x};\boldsymbol{\theta}) = \frac{\sum_{\mathbf{h}} \exp \Psi(y, \mathbf{h}, \mathbf{x};\boldsymbol{\theta})}{\sum_{y', \mathbf{h}} \exp \Psi(y', \mathbf{h}, \mathbf{x};\boldsymbol{\theta})}\,, \tag{1}$$

where $\mathbf{h}$ is a set of hidden variables $\mathbf{h}^k = \{h_1, h_2, \ldots, h_T\}$, with $\mathbf{h}^k \in \mathcal{H}$, and $\Psi(\cdot)$ is the potential function that specifies dependencies between different nodes in the model given by:

$$\Psi(y, \mathbf{h}, \mathbf{x};\boldsymbol{\theta}) = \sum_{j\in\mathcal{V}} \boldsymbol{\theta}_1 \cdot \psi_1(x_j, h_j) + \sum_{j\in\mathcal{V}} \boldsymbol{\theta}_2 \cdot \psi_2(y, h_j) + \sum_{(i,j)\in\mathcal{E}} \boldsymbol{\theta}_3 \cdot \psi_3(y, h_i, h_j)\,. \tag{2}$$

The functions $\boldsymbol{\theta}_1 \cdot \psi_1(\cdot)$ and $\boldsymbol{\theta}_2 \cdot \psi_2(\cdot)$ are the node potentials N_p, which model the relationship between the hidden variable h_j and the feature vector x_j, and the relationship between the class label y and the hidden variable h_j, respectively, and $\boldsymbol{\theta}_3 \cdot \psi_3(\cdot)$ is the edge potential E_p, which models the relationship between the class label y and the hidden variables h_i and h_j.

3 Estimation of the Number of Hidden States

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

For a given label $y = \alpha$, the node potentials for all observations and all possible hidden variables can be represented in a matrix form:

$$N_p = [n_{p_{ij}}]_{S\times T} = \begin{bmatrix} \theta_{1_{11}} \cdot x_1 + \theta_{2_{1\alpha}} & \ldots & \theta_{1_{1T}} \cdot x_T + \theta_{2_{1\alpha}} \\ \theta_{1_{21}} \cdot x_1 + \theta_{2_{2a}} & \ldots & \theta_{1_{2T}} \cdot x_T + \theta_{2_{2\alpha}} \\ \vdots & \ddots & \vdots \\ \theta_{1_{S1}} \cdot x_1 + \theta_{2_{S\alpha}} & \ldots & \theta_{1_{ST}} \cdot x_T + \theta_{2_{T\alpha}} \end{bmatrix}, \tag{3}$$

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

$$E_p = [e_{p_{ij}}]_{S \times S} = \begin{bmatrix} \theta_{3_{11}} & \theta_{3_{12}} & \dots & \theta_{3_{1S}} \\ \theta_{3_{21}} & \theta_{3_{22}} & \dots & \theta_{3_{2S}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{3_{S1}} & \theta_{3_{S2}} & \dots & \theta_{3_{SS}} \end{bmatrix}. \tag{4}$$

All values of the node potentials N_p matrix are transformed into the range $[0, 1]$ using min-max normalization, so that all input variables equally contribute in the model and the parameters of the node potential N_p are not scaled with respect to the units of the inputs. As a result, we end up with smaller standard deviations, which can suppress the effect of outliers. Also, we construct a stochastic matrix based on the edge potential E_p, with each row summing to one.

To determine the number of hidden variables, we employ multiple HMMs. Specifically, an HMM is defined by the set of hidden variables $\mathbf{h}$ and a set of parameters $\boldsymbol{\Lambda} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$, where $\boldsymbol{\pi}$ is a vector that collects the prior probabilities of h_i being the first hidden variable of a state sequence, $\mathbf{A}$ is a matrix that collects the transition probabilities of moving from one hidden variable to another, and $\mathbf{B}$ is the matrix that collects the emission probabilities, which characterize the likelihood of a certain observation $\mathbf{x}$, if the model is in hidden variable h_i.

Let us consider that the normalized node potentials are the entries of the emission probability matrix, the edge potentials are the entries of the transition probability matrix, and there is a vector of initial probabilities $\boldsymbol{\pi}$, where each hidden variable is equally probable to be first in the sequence $\pi_{1 \times S} = \frac{1}{S} \cdot \mathbb{1}_S$.

Given the above definitions, for a given label $y = \alpha$, we can determine an HMM and the optimal hidden variable sequence using the Viterbi algorithm to estimate the probability of the most probable path ending $\delta_t(i)$ and keep track of the best path ending $\phi_t(i)$ in the i^{th} hidden state at time t:

$$\delta_t(i) = \max_{h_1, \dots, h_{T-1}} P(h_1, \dots, h_{T-1}, h_T = i, x_1, \dots, x_T | \pi, E_p, N_p), \tag{5}$$

$$\phi_t(i) = \operatorname*{argmax}_{h_1, \dots, h_{T-1}} P(h_1, \dots, h_{T-1}, h_T = i, x_1, \dots, x_T | \pi, E_p, N_p). \tag{6}$$

Figure 1 depicts the flow of the proposed method for the estimation of the optimal number of hidden variables. The proposed method learns the optimal number of hidden variables following an incremental learning approach. It starts by setting an initial number $S = |\mathcal{H}| \geq 1$ for the hidden variables and the maximum number of iterations. In each iteration, all optimal paths, for every video sequence and for every label, are estimated using the Viterbi algorithm and the frequency of appearance of each hidden state in all paths is calculated. The termination criterion is reached when the frequency of each hidden variable is lower than a predefined threshold τ. If this criterion is not satisfied, the number

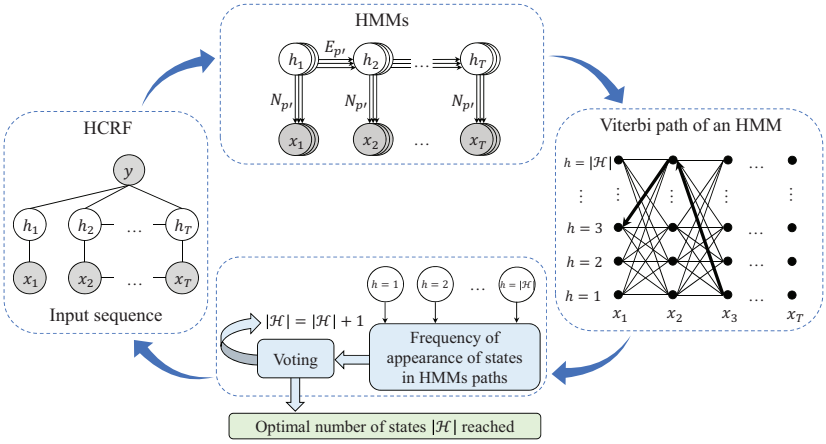


Fig. 1. Illustration of the iterative and incremental method for the estimation of the optimal number of the hidden variables. The grey nodes are the observed features (x_i), and the unknown labels (y). The white nodes are the hidden variables (h). At each iteration, an HCRF is built with an initial number of hidden states and then the Viterbi algorithm is used to estimate the optimal paths in the HMM setting. Then, the frequency of each hidden variable is computed and if the termination criterion is satisfied the number of hidden states is decided by majority voting, otherwise the number of hidden states is increased by one and the process is repeated.

of hidden variables is increased by one and the process is repeated. If the termination criterion is satisfied, we move to the next iteration and a voting for the most probable number of hidden variables in the current iteration is performed. Finally, when the maximum number of iterations is reached, the optimal number of hidden variables is the one with the majority of votes.

4 A Student's t-mixture Prior on the Model Parameters

Let us assume that the parameters $\boldsymbol{\theta}$ of the proposed RI-HCRF follow a mixture model with three Student's-t components. Taking into consideration that the parameter vector $\boldsymbol{\theta}$ describes three different relationships among observations, hidden variables and labels, we expect that each component corresponds to one of these relationships. To this end, the proposed method relies on partitioning the parameter vector using a Student's t-mixture model to identify, preserve, and enhance the different characteristic of each partition and improve the classification model. The use of Student's t-mixture model is justified by the fact that it has a heavier tails pdf and provides smaller weights to the observations that lie in the tail area. Thus, it provides robustness to outliers and less extreme estimates of the posterior probabilities of the mixture model [11]. Additionally, each Student's t-component originates from a wider class of an elliptically symmetric distribution with an additional robustness tuning parameter that corresponds to the degrees of freedom ν.

However, by making the above assumption the problem of setting the mixture weights arises. To estimate the best weights for the mixture model, one might perform an exhaustive search to check multiple combinations of probable values of mixture weights. To avoid the prohibitive quadratic computational cost of this approach, we dynamically estimate the best fitted mixture model for parameters $\boldsymbol{\theta}$ at the end of each iteration of the training process.

Let each parameter $\boldsymbol{\theta}$ follows a univariate t-distribution with mean μ, variance σ^2, and $\nu \in [0, \infty)$ degrees of freedom then, given the weight u that follows a Gamma distribution (Γ) parameterized by ν, the parameter $\boldsymbol{\theta}$ has the univariate normal with mean μ and variance σ^2/u, with u being a random variable distributed as $u \sim \Gamma(\nu/2, \nu/2)$.

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

$$p(\boldsymbol{\theta}; \nu, \mu, \lambda) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \Big(\frac{\lambda}{\pi\nu}\Big)^{\frac{1}{2}} \Big(1 + \frac{\lambda(\theta-\mu)^2}{\nu}\Big)^{-\frac{\nu+1}{2}}, \tag{7}$$

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

$$\phi(\boldsymbol{\theta}, \Omega) = \sum_{i=1}^{K} \pi_i p(\theta; \nu_i, \mu_i, \lambda_i), \tag{8}$$

where $\boldsymbol{\theta}$ denotes the observed data vector, $\Omega = \{\Omega_i\}_{i=1}^K$ is the mixture parameter set with $\Omega_i = \{\pi_i, \nu_i, \mu_i, \lambda_i\}$, and π_i are the i^{th} mixing proportions that satisfy the following constraints: $\sum_{i=1}^K \pi_i = 1$ and $0 \leq \pi_i \leq 1$. The best fitted mixture with Student's t-components can be obtained by maximizing the likelihood function using the EM algorithm [11].

During training a maximum likelihood approach is followed to estimate the parameters $\boldsymbol{\theta}$ of the model by maximizing the following loss function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(y_i | x_i; \boldsymbol{\theta}) + \log\Big(\sum_{k=1}^{K} \pi_k p(\boldsymbol{\theta}; \nu_k, \mu_k, \lambda_k)\Big), \tag{9}$$

where the first term is the conditional log-likelihood of the input data and the second term represents the best fitted Student'st-mixture model on parameter vector $\boldsymbol{\theta}$, obtained by the EM algorithm. The optimal weights $\boldsymbol{\theta}^\star$ are learned by maximizing the objective function. The optimization of Eq. (9) is performed using the limited-memory BFGS (LBFGS) method [4], since the value and the derivative of the objective function may be calculated. Then, the corresponding label is estimated by maximizing the posterior probability:

$$y^\star = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} P(y | \mathbf{x}; \boldsymbol{\theta}^\star). \tag{10}$$

Table 1. Comparison of the classification accuracies (%) for the Parliament [20], TVHI [12], and SBU [24] datasets. The results were averaged for all different configurations (mean ± standard deviation).

Method	Parliament [20]	TVHI [12]	SBU [24]
Hand-crafted features			
SVM [6]	75.5 ± 3.1	85.5 ± 4.8	81.5 ± 6.8
CRF [9]	82.1 ± 1.7	$\mathbf{100.0} \pm 0.0$	87.3 ± 4.4
HCRF [13]	82.9 ± 2.5	99.5 ± 1.5	87.0 ± 2.6
RI-HCRF	$\mathbf{85.5} \pm 3.1$	$\mathbf{100.0} \pm 0.0$	$\mathbf{89.4} \pm 2.7$
CNN-based features			
SVM [6]	73.3 ± 0.9	91.0 ± 0.6	94.3 ± 0.4
CRF [9]	85.5 ± 0.7	92.8 ± 0.3	93.7 ± 0.4
HCRF [13]	89.0 ± 0.8	93.0 ± 0.2	91.5 ± 0.4
CNN [18]	78.1 ± 0.4	60.5 ± 1.1	94.2 ± 0.8
RI-HCRF	$\mathbf{89.5} \pm 0.7$	$\mathbf{93.5} \pm 0.2$	$\mathbf{94.8} \pm 0.3$

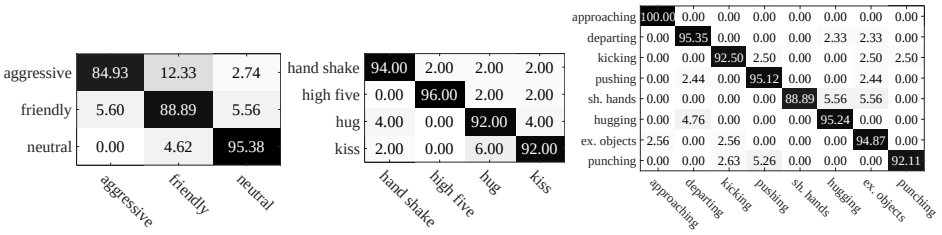
5 Experimental Results

To demonstrate the ability of the proposed method to recognize human actions, we compared it with several state-of-the-art methods in three publicly available benchmark datasets. First, we used the Parliament dataset [20], which consists of 228 video sequences of political speeches categorized in three behavioral classes: friendly, aggressive, and neutral. The TV human interaction (TVHI) dataset [12], is a collection of 300 video sequences depicting four kinds of interactions such as handshakes, high fives, hugs, and kisses. Finally, the SBU Kinect Interaction (SBU) dataset [24] is used, which contains approximately 300 video sequences of two-person interactions captured by a Microsoft Kinect sensor. Each video is labeled with one of the following actions: approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands.

For all datasets, we used spatio-temporal interest points (STIP) [10] as our basic representation. Also, for the Parliament and TVHI datasets, we extracted the mel-frequency cepstral coefficients (MFCC) [14] features along with their first and second order derivatives, while for the TPI dataset, we used the provided by the dataset poses. Additional to the hand-crafted features, we used CNNs both for end-to-end classification and for feature extraction by employing the pre-trained model of Tran *et al.* [18], which is a 3D ConvNet. Finally, we assessed the performance of the RI-HCRF by comparing with the following baseline methods: SVM [6], CRF [9], HCRF [13], and CNN (end-to-end) [18].

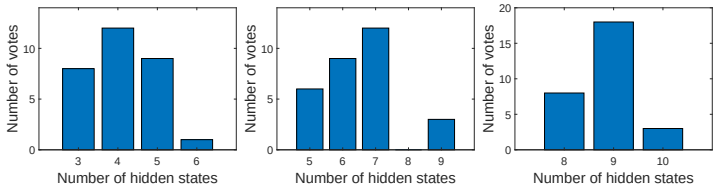
The threshold for the automatic learning of the optimal number of hidden variables was set to take values from a discrete set $\tau \in \{0.001, 0.005, 0.01, 0.02, 0.05\}$ and the maximum number of iterations were set to 30. The number of components for the mixture of Student's t-distribution was set to $K = 3$. The model parameters were randomly initialized and the experiments were repeated five times, while 5-fold cross validation was used to split into training and test sets. To examine the performance of the RI-HCRF model against the standard HCRF, we varied the number of hidden variables from 3 to 18.

The average recognition accuracies and the corresponding standard deviations for all datasets, for both hand-crafted and CNN features, are presented in



(a) (b) (c)

Fig. 2. Confusion matrices for the classification results using CNN features for the (a) Parliament [20], (b) TVHI [12], and (c) SBU [24] datasets.



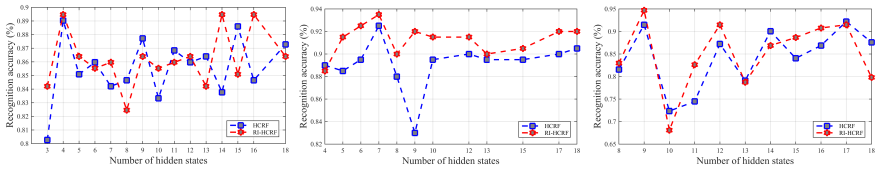
(a) (b) (c)

Fig. 3. Estimation of the optimal number of hidden variables by the proposed RI-HCRF algorithm using the CNN features for the (a) Parliament [20], (b) TVHI [12], and (c) SBU [24] datasets. The number of hidden states that do not appear in the horizontal axis received zero votes.

Table 1. It can be observed that the proposed RI-HCRF method outperforms all the state-of-the-art methods for all datasets. It is worth noting that RI-HCRF show very small deviation compared with the rest of the methods, which indicates that the use of Student's t-distribution provides robustness and reduces the miss classification errors. It is also interesting to observe that for the Parliament and TVHI datasets, the absolute improvement of RI-HCRF over the CNN model is very high (11% and 33%, respectively). This improvement can be explained by the fact that the CNN model uses a linear classifier in the softmax layer, while RI-HCRF is more suitable to encode sequential data by modeling dependencies between consecutive frames in a more principled way.

Since the best results, for the RI-HCRF method, were achieved when CNN features were employed, the corresponding confusion matrices are depicted in Fig. 2. It can be observed that the misclassification errors are quite small for the three datasets indicating that the proposed method can efficiently recognize human actions with high accuracy and small intra-class classification errors.

Figure 3 depicts the results for the prediction of the optimal number of hidden states for all three datasets, when CNN features are used for the classification. For the Parliament dataset, number 4 is the most suitable candidate, for TVHI number 7 seems to be the most probable case, and for the SBU dataset, number 9 turns out to be the candidate with the most votes estimated by the RI-HCRF model. The estimated number of hidden states obtained from the proposed model



(a) (b) (c)

Fig. 4. Classification accuracies with respect to the number of hidden states using exhaustive search for the (a) Parliament [20], (b) TVHI [12], and (c) SBU [24] datasets. The results from exhaustive search are in-line with the optimal number of hidden states as predicted by RI-HCRF in Fig. 3.

is in fully agreement with the results from the exhaustive search of the number of hidden states (Fig. 4). Also, we may observe that the number of candidates is much lower compared to the exhaustive search.

6 Conclusion

In this paper, a video-based action recognition method using a graphical representation of human actions is introduced. The proposed approach is an extension of standard HRCFs, which can automatically infer the number of hidden states from the input data. The proposed model is a mixture of three Student's t-components coupled to the RI-HCRF model as a prior to the parameters providing robustness to outliers. An extented experimental evaluation demonstrated that the proposed method achieved promising results, while reduced the search space for the estimation of the number of hidden states.

Acknowledgments: This work has been funded by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund. 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.

References

1. Bouchard, G.: Bias-variance tradeoff in hybrid generative-discriminative models. In: ICMLA. pp. 124–129 (2007)
2. Bousmalis, K., Zafeiriou, S., Morency, L.P., Pantic, M.: Infinite hidden conditional random fields for human behavior analysis. Transactions on Neural Networks and Learning Systems **24**(1), 170–177 (2013)
3. Bousmalis, K., Zafeiriou, S., Morency, L.P., Pantic, M., Ghahramani, Z.: Variational hidden conditional random fields with coupled dirichlet process mixtures. In: Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 531–547 (2013)

4. Byrd, R.H., Nocedal, J., Schnabel, R.B.: Representations of quasi-Newton matrices and their use in limited memory methods. Mathematical Programming **63**(1), 129–156 (1994)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (July 2017)
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011)
7. 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: CVPR. pp. 2625–2634 (June 2015)
8. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (June 2016)
9. Lafferty, J.D., andF. C. N. Pereira, A.M.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. pp. 282–289 (2001)
10. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2-3), 107–123 (2005)
11. McLachlan, G.J., Peel, D.: Robust mixture modelling using the t distribution. Statistics and Computing **10**(4), 335–344 (2000)
12. 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)
13. 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)
14. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
15. Sigurdsson, G.A., Russakovsky, O., Gupta, A.: What actions are needed for understanding human actions in videos? In: Proc. IEEE International Conference on Computer Vision. pp. 2156–2165 (2017)
16. Song, Y., Morency, L.P., Davis, R.: Multi-view latent variable discriminative models for action recognition. In: CVPR. Providence, RI (June 2012)
17. Soullard, Y., Artières, T.: Hybrid HMM and HCRF model for sequence classification. In: ESANN. pp. 453–458 (2011)
18. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV. pp. 4489–4497 (2015)
19. Vrigkas, M., Kazakos, E., Nikou, C., Kakadiaris, I.A.: Inferring human activities using robust privileged probabilistic learning. In: ICCVW. pp. 2658–2665 (2017)
20. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: Classifying behavioral attributes using conditional random fields. In: SETN. vol. 8445, pp. 95–104 (May 2014)
21. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and AI **2**(28), 1–26 (2015)
22. 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)
23. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR. pp. 4325–4334 (July 2017)
24. Yun, K., Honorio, J., Chattopadhyay, D., Berg, T.L., Samaras, D.: Two-person interaction detection using body-pose features and multiple instance learning. In: CVPRW. pp. 28–35 (2012)
25. Zhang, J., Gong, S.: Action categorization with modified hidden conditional random field. Pattern Recognition **43**(1), 197–203 (2010)