MERGING LINEAR DISCRIMINANT ANALYSIS WITH BAG OF WORDS MODEL FOR HUMAN ACTION RECOGNITION

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: {tefas,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper we propose a novel method for human action recognition, that unifies discriminative Bag of Words (BoW)-based video representation and discriminant subspace learning. An iterative optimization scheme is proposed for sequential discriminant BoWs-based action representation and codebook adaptation based on action discrimination in a reduced dimensionality feature space where action classes are better discriminated. Experiments on four publicly available action recognition data sets demonstrate that the proposed unified approach increases the discriminative ability of the obtained video representation, providing enhanced action classification performance.

Index Terms— Bag of Words, Discriminant Learning

1. INTRODUCTION

Human action recognition from videos has received considerable attention in the last two decades due to its importance in a wide range of applications, like movie (post-)processing and human-computer interaction (HCI). It is, still, an active research field due to its difficulty, which is, mainly, caused because there is not a formal description of actions. Action execution style variations and changes in human body sizes among individuals, as well as different camera observation angles are some of the reasons that lead to high intra-class and, possibly, small inter-class variations of action classes.

The state-of-the-art approach to date involves two processing steps, i.e., video representation and classification. In the first processing step, a vectorial video representation highlighting the properties of the depicted action and (possibly) discriminating the properties of different action classes is employed. Recently, several action descriptors aiming at action recognition in unconstrained environments have been proposed, including local sparse and dense space-time features [1, 2, 3, 4]. Such descriptors capture information appearing in video frame locations that either correspond to video frame interest points which are tracked during action execution, or that are subject to abrupt intensity value variations and, thus, contain information regarding motion speed and/or acceleration, which is of interest for the description of actions. These local video frame descriptors are calculated by using the color (grayscale) video frames and, thus, video frame segmentation is not required.

After describing actions, videos depicting actions, called action videos hereafter, are usually represented by fixed size vectors. Perhaps the most well studied and successful approach for action representation is based on the Bag of Words (BoWs) model [5], in which each video is represented by a vector obtained by applying (hard or soft) quantization on its descriptors using a set of descriptor prototypes forming the so-called codebook. This codebook is determined by clustering the features describing training action videos. The BoWs-based action representation has been combined with several classifiers, like Support Vector Machines, Artificial Neural Networks and Discriminant Analysis based classification schemes, providing high action classification performance on publicly available data sets aiming at different application scenarios.

In this paper, we build on the BoWs-based video representation by introducing discriminative criteria on the codebook learning process. Contrary to the usual approach, where the video representation and classification steps are performed independently, the proposed method integrates video representation and classification in a multi-class optimization process in order to produce a discriminant BoWs-based video representation and an optimized classification scheme. Two processing steps are iteratively repeated to this end. The first one, involves the calculation of BoWs-based video representations of increased discrimination power, while the second exploits these video representations in order to learn a classification scheme involving optimal data projection to a discriminant subspace. By following this approach, the proposed method is able to learn a BoWs-based video representation enhancing action discrimination and, thus, achieve better classification performance in a wide range of action recognition problems.

The rest of the paper is structured as follows. Related work is discussed in Section 2. The proposed method for integrated discriminant BoWs-based video representation learning is described in Section 3. Experiments conducted on pub-

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART).

licly available data sets aiming at different application scenarios are illustrated in Section 4. Finally, conclusions are drawn in Section 5.

2. RELATED WORK

Let us denote by $\mathcal{U}$ a video database containing N_T videos followed by action class labels l_i, $i = 1, \dots, N_T$ appearing in an action class set $\mathcal{A} = \{\alpha\}_{\alpha=1}^{C}$. Let us assume that for each video i a set of N_i descriptors $\mathbf{p}_{ij}, \in \mathbb{R}^D$, $i = 1, \dots, N_T$, $j = 1, \dots, N_i$ have been calculated, which are normalized in order to have unit l_2 norm.

Standard BoW-based video representation, apples a clustering technique, e.g. K-Means, on the descriptors $\mathbf{p}_{ij}$ calculated for all the N_T training videos without exploiting the labeling information that is available for the training videos, in order to determine K codebook vectors $\mathbf{v}_k \in \mathbb{R}^D$ forming the so-called codebook $\mathbf{V} \in \mathbb{R}^{D \times K}$. After determining the codebook $\mathbf{V}$, the representation of video i is obtained by applying hard or soft vector quantization on the descriptors $\mathbf{p}_{ij}$, $j = 1, \dots, N_i$. In the first case, the BoWs-based representation of action video i is a histogram of features, calculated by assigning each feature vector $\mathbf{p}_{ij}$ to the cluster of the closest codebook vector $\mathbf{v}_k$. In the second case, a distance function, usually the Euclidean one, is used in order to determine N_i distance vectors, each denoting the similarity of feature vector $\mathbf{p}_{ij}$ to all the codebook vectors $\mathbf{v}_k$, and the representation of action video i is determined to be the mean normalized distance vector [6].

The above-described BoW-based video representation has been shown to provide satisfactory performance in many action recognition problems. However, due its unsupervised nature, the discriminative ability of the BoWs-based action representation is limited. In order to increase the quality of the adopted codebook, codebook adaptation processes have been proposed which adopt a generative approach. That is, the initial codebook generated by clustering the features describing training videos is adapted so as to reduce the reconstruction error of the resulted video representation [7]. However, since this generative adaptation process does not take into account the class labels that are available for the training action videos, the discriminative ability of the optimized codebook is not necessarily increased. In order to increase the discriminative ability of the adopted codebook, discriminative codebook learning processes [8, 9] have been proposed. However, since the codebook calculation process is, still, disconnected from the adopted classification scheme, the obtained codebook may not be the one that is best suited for the task under consideration, i.e., the classification of actions in our case.

A method aiming at simultaneously learning both a discriminative codebook and a classifier is proposed in [10] for image classification. It consists of two iteratively repeated steps. The first one involves training images representation by a set of class-specific histograms of visual words at the bit level and multiple binary classifiers, one for each image category, training by using the obtained histograms. While this approach has lead to increased image classification performance, its extension in other classification tasks, e.g., action recognition, is not straightforward. Another approach has been proposed in [11], where a two-class linear SVM-based codebook adaptation scheme is formulated. The adoption of a two class formulations (adopted in both [10, 11]) generates the drawback that $C(C-1)/2$ two-class codebooks have to be learned and used in the test phase along with an appropriate fusion strategy. In addition, such an approach is not able to exploit inter-class correlation information appearing in multi-class problems, which may facilitate class discrimination. The proposed method, by employing a multi-class learning approach overcomes these drawback.

3. PROPOSED METHOD

The proposed method exploits a generalization of the Euclidean distance, i.e., $d_{ijk} = \|\mathbf{v}_k - \mathbf{p}_{ij}\|_2^{-g}$, in order to define the similarity between descriptor $\mathbf{p}_{ij}$ and the codebook vector $\mathbf{v}_k$. The parameter g is used in order to define the type of the adopted quantization, i.e. a value $g = 1.0$ leads to a BoWs-based representation using soft vector quantization, while a value $g \gg 1.0$ leads to a BoWs-based representation using hard vector quantization. Membership vectors $\mathbf{u}_{ij} \in \mathbb{R}^K$, encoding the similarity of $\mathbf{p}_{ij}$ to all the codebook vectors $\mathbf{v}_k$, are obtained by normalizing the distance vectors $\mathbf{d}_{ij} = [d_{ij1} \dots d_{ijK}]^T$ in order to have unit l_1 norm, i.e. $\mathbf{u}_{ij} = \mathbf{d}_{ij}/\|\mathbf{d}_{ij}\|_1$. The BoWs-based representation of action video i is obtained by calculating the mean membership vector $\mathbf{q}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{u}_{ij}$. Finally, the mean membership vectors $\mathbf{q}_i$ are normalized in order to produce the so-called action vectors $\mathbf{s}_i = \mathbf{q}_i/\|\mathbf{q}_i\|_2$. After calculating the action vectors representing all the action videos, they are normalized in order to have zero mean and unit standard deviation, resulting to the normalized action vectors $\mathbf{x}_i \in \mathbb{R}^K$. In order to map the normalized action vectors $\mathbf{x}_i$ to a new feature space in which action classes are better discriminated, an optimal linear transformation $\mathbf{W}^*$ is obtained by solving the *trace ratio* [12, 13] optimization problem used in Linear Discriminant Analysis (LDA) [14], i.e.:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \frac{trace\{\mathbf{W}^T \mathbf{S}_w \mathbf{W}\}}{trace\{\mathbf{W}^T \mathbf{S}_b \mathbf{W}\}}, \qquad (1)$$

where $\mathbf{S}_w$, $\mathbf{S}_b$ are within-class and between-class scatter matrices calculated using the training action vectors $\mathbf{s}_i$, $i = 1, \dots, N_T$ [15]. Finally, the discriminant action vectors $\mathbf{z}_i$ are obtained by applying $\mathbf{z}_i = \mathbf{W}^{*T}\mathbf{x}_i$.

After determining the above described discriminant action video representation, a codebook adaptation process is performed in order to increase the codebook discriminative ability based on action class discrimination in the obtained discriminant space. The procedure followed to this end is described in the following.

3.1. Codebook Adaptation

Since the normalized action vectors $\mathbf{x}_i$ are functions of the adopted codebook $\mathbf{V}$, the optimization problem (1) is a function of both the projection matrix $\mathbf{W}$ and the codebook $\mathbf{V}$. Based on this observation, we propose to minimize the trace ratio criterion with respect to both $\mathbf{W}$ and $\mathbf{V}$, in order to simultaneously increase the codebook discriminative ability and to obtain the optimal transformation matrix for action classes discrimination:

$$\mathcal{J}(\mathbf{W},\mathbf{V}) = \frac{trace\{\mathbf{W}^T\mathbf{S}_w(\mathbf{V})\mathbf{W}\}}{trace\{\mathbf{W}^T\mathbf{S}_b(\mathbf{V})\mathbf{W}\}} \qquad (2)$$

We propose an iterative optimization scheme to this end formed by two steps:

- For a given codebook $\mathbf{V}_t$, training normalized action vectors $\mathbf{x}_{i,t}$ are employed in order to determine the optimal projection matrix $\mathbf{W}_t^*$ by solving the trace ratio problem (2).

- Codebook vectors $\mathbf{v}_{k,t}$ are adapted, in the direction of the gradient of (2), by using the obtained $\mathbf{W}_t^*$. The adaptation of $\mathbf{v}_{k,t}$ is performed by following the gradient of $\mathcal{J}$ with respect to $\mathbf{v}_{k,t}$:

$$\mathbf{v}_{k,t+1} = \mathbf{v}_{k,t} - \eta \frac{\partial \mathcal{J}_t}{\partial \mathbf{v}_{k,t}}, \qquad (3)$$

$$\begin{aligned}\frac{\partial \mathcal{J}_t}{\partial \mathbf{v}_{k,t}} &= \left(a\tilde{\mathbf{W}}_{t(i,:)}(\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t^{\alpha}) - c\tilde{\mathbf{W}}_{t(i,:)}\bar{\mathbf{x}}_t^{\alpha})\right) \\ \cdot &\quad \left(\frac{1}{\tilde{s}_{k,t}} - \frac{s_{ik,t} - \bar{s}_{k,t}}{\tilde{s}_{k,t}^3}\right)\left(\frac{1}{\|\mathbf{q}_{i,t}\|_2} - \frac{q_{ik,t}^2}{\|\mathbf{q}_{i,t}\|_2^3}\right) \\ \cdot &\quad \frac{N_T - 1}{N_T N_i}\left(\frac{1}{\|\mathbf{d}_{ij,t}\|_1} - \frac{d_{ijk,t}}{\|\mathbf{d}_{ij,t}\|_1^2}\right) \\ \cdot &\quad -g\|\mathbf{v}_{k,t} - \mathbf{p}_{ij}\|_2^{-(g+2)}\left(\mathbf{v}_{k,t} - \mathbf{p}_{ij}\right), \qquad (4)\end{aligned}$$

where η is an update rate parameter. In order to avoid scaling issues, codebook vectors of both the initial and the updated codebooks, $\mathbf{v}_{k,0}$, $\mathbf{v}_{k,t}$ respectively, are normalized to have unit l_2 norm.

In order to accelerate the codebook adaptation process and (possibly) to avoid convergence on local minima, in our experiments we have employed a (dynamic) line search strategy, where in each iteration of the codebook adaptation process, the trace ratio criterion (2) was evaluated by using (3) and $\eta_0 = 0.1$. In the case where $\mathcal{J}_{t+1} < \mathcal{J}_t$, the trace ratio criterion was evaluated by using a codebook update parameter value $\eta_n = 2\eta_{n-1}$. This process is followed until $\mathcal{J}_{t+1} > \mathcal{J}_t$ and the codebook update parameter value providing the highest $\mathcal{J}$ decrease was employed for codebook adaptation. In the case where, by using a codebook update parameter value $\eta_0 = 0.1$, $\mathcal{J}_{t+1} > \mathcal{J}_t$, the trace ratio criterion was evaluated by using a codebook update parameter value $\eta_n = \eta_{n-1}/2$. This process is followed until $\mathcal{J}_{t+1} < \mathcal{J}_t$ and the codebook update parameter value providing $\mathcal{J}$ decrease was employed for codebook adaptation.

3.2. Action Recognition (Test Phase)

Let us denote by $\mathbf{V}_{opt}$, $\mathbf{W}_{opt}$ the codebook and the corresponding projection matrix obtained by applying the above described optimization process employing the feature vectors $\mathbf{p}_{ij}$ describing the training action videos and the corresponding action class labels. Let $\mathbf{p}_{tj} \in \mathbb{R}^D$, $j = 1, \dots, N_t$ be feature vectors describing a test action video. $\mathbf{p}_{tj}$ are employed in order to calculate the corresponding normalized action vector $\mathbf{x}_t \in \mathbb{R}^K$ using $\mathbf{V}_{opt}$. $\mathbf{x}_t$ can be either classified in this space, or be mapped to the discriminant space, determined in the training phase, by applying $\mathbf{z}_t = \mathbf{W}_{opt}^T\mathbf{x}_t$. In either cases, action classification is performed by employing any, linear or non-linear, classifier, like K-NN, SVM and ANNs.

4. EXPERIMENTAL EVALUATION

In this Section we present experiments conducted in order to evaluate the proposed discriminant BoWs-based action representation. In all the experiments we have employed the Harris3D detector [16] followed by HOG/HOF descriptors [1] calculation for video description. The optimal values of parameters K and g have been determined by applying grid search using values $50 < K < 500$ and $g = [1, 2, 5, 10, 20]$, respectively. In order to limit the complexity, we cluster a subset of $100k$ randomly selected HOG/HOF descriptors for initial codebook calculation. To increase precision of the initial codebook, we initialize K-Means 10 times and keep the codebook providing the smallest error. In the test phase, classification is performed by employing a Single-hidden Layer Feedforward Neural Network trained by applying the recently proposed Extreme Learning Machine algorithm ([17]).

We have used four publicly available data sets aiming at different application scenarios, i.e., the KTH [18], the Hollywood2 [19], the Ballet [20] and the i3DPost [21] data sets. We have adopted the experimental protocols suggested by the databases. In all the experiments we compare the performance of the standard BoWs-based video representation (BoWs) and the proposed discriminant BoWs-based representation (DBoWs). Furthermore, we provide comparison results of the proposed discriminant BoWs-based video representation adopting the above mentioned descriptor-classifier combination with some recently proposed state-of-the-art action recognition methods evaluating their performance on the adopted action recognition data sets.

Table 1. Classification rates on the KTH data set.

	Representation	Performance
Method [25]	low-level	82%
Method [26]	low-level	84.3%
Method [27]	low-level	87.3%
Method [28]	low-level	90.57%
Method [29]	low-level	91.1%
Method [22]	high-level	94.5%
Method [23]	high-level	98.9%
Method [24]	**high-level**	**99.54%**
BoWs	low-level	88.89%
DBoWs	**low-level**	**92.13%**

Table 2. Classification rates on the Hollywood2 data set.

	Representation	Performance
Method [19]	Harris3D+HOG+HOF	32.4%
Method [19]	Harris3D+HOG+HOF+SIFT	32.6%
Method [19]	Harris3D+HOG+HOF+SIFT+Scene	35.5%
Method [1]	Harris3D+HOG/HOF	**45.2%**
Method [4]	Dense+HOG	41.5%
Method [4]	Dense+HOF	50.8%
Method [1]	Dense+HOG/HOF	47.4%
Method [4]	Dense+HOG+HOF+MBH+Traj	**58.3%**
Method [30]	Regions+HOG+HOF+OF	41.34%
BoWs	Harris3D+HOG/HOF	41.5%
DBoWs	Harris3D+HOG/HOF	**45.8%**

Table 3. Classification rates on the Ballet data set.

Method [3]	Method [31]	BoWs	**DBoWs**
91.1%	**91.3%**	86.3%	**91.1%**

Table 4. Classification rates on the i3DPost data set.

Method [6]	Method [32]	BoWs	**DBoWs**
94.87%	**98.44%**	95.31%	**98.44%**

4.1. Experimental Results

Tables 1 - 4 illustrate the performance obtained by using the proposed discriminant BoWs-based video representation on the KTH, Hollywood2, Ballet and i3DPost data sets, respectively. As can be seen in these Tables, the adoption of an approach integrating the video representation and classification steps enhances performance, when compared to the standard approach where these steps are performed independently, since the proposed method consistently provides better performance. In Tables 1 - 4 we, also, compare the performance of the proposed action video recognition approach with that of some state-of-the-art methods, recently proposed in the literature.

On the KTH data set, the use of the BoWs-based action video representation led to a classification rate equal to 88.89%. By adopting the proposed DBoWs-based action video representation, an increased classification rate, equal to 92.13%, has been obtained. The proposed method outperforms other state-of-the-art methods employing low-level video representations, while it provides performance comparable with the method in [22] and inferior performance when compared with the methods in [23], [24], which exploit high-level video representations. However, the calculation of high-level representations is computationally demanding, compared to the calculation of low-level ones and, thus, a comparison between the two approaches in terms of only the obtained action recognition performance is not fair.

On the Hollywood2 data set, the use of the BoWs-based action video representation led to a performance equal to 41.5%. By adopting the proposed DBoWs-based action video representation, a performance equal to 45.8% has been obtained. The methods evaluated on this data set can be roughly divided based on the employed action video description. Methods employing densely sampled descriptors for action video representation have been shown to outperform the ones employing descriptors calculated on STIPs. Taking into account that STIP-based video representations are much faster, when compared to ones exploiting densely-sampled visual information, we can see that a comparison between the two approaches, in terms of only the obtained performance, is not fair. It can also be seen that the adoption of a higher number of descriptors, each describing a different action property, enhances action classification performance.

On the Ballet data set, the use of the BoWs-based action video representation led to an action classification rate equal to 86.3%. The use of the proposed DBoWs-based action video representation increased the action classification rate to 91.1%, which is comparable to the performance of the two competing methods presented in Table 3.

Finally, on the i3DPost data set, the use of the BoWs-based action representation led to an action classification rate equal to 95.31%, while the adoption of the proposed DBoWs-based action video representation led to an action classification rate equal to 98.44%, equal to that of [32], which employs a computationally expensive $4D$ optical flow-based action video representation and, thus, its operation is slower, when compared with the proposed method in the test phase.

5. CONCLUSIONS

In this paper we proposed a novel human action video classification method, which unifies discriminative codebook calculation and discriminant subspace learning. An iterative optimization scheme has been proposed for sequential discriminant BoWs-based action video representation calculation and codebook adaptation based on action classes discrimination. Experiments conducted on four publicly available action recognition data sets aiming at different application scenarios show that the proposed unified approach increases the codebook discriminative ability providing enhanced action video classification performance, since it consistently outperforms the standard approach, where video representation and classification are performed independently.

6. REFERENCES

[1] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” *British Machine Vision Conference*, pp. 1–11, 2009.

[2] L. Yeffet and L. Wolf, “Local trinary patterns for human action recognition,” *International Conference on Computer Vision*, pp. 492–497, 2009.

[3] T. Guha and R.K. Ward, “Learning sparse representations for human action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2011.

[4] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, “Action recognition by dense trajectories,” *Computer Vision and Pattern Recognition*, 2011.

[5] G. Csurka, C. Bray, C. Dance, and L. Fan, “Visual categorization with bags of keypoints,” *European Conference on Computer Vision*, 2004.

[6] A. Iosifidis, A. Tefas, and I. Pitas, “View-invariant action recognition based on artificial neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–425, 2012.

[7] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classication,” *Computer Vision and Pattern Recognition*, 2010.

[8] J. Winn, A. Criminisi, and T. Minka, “Object categorizations by learned universal visual dictionary,” *International Conference on Computer Vision*, 2005.

[9] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, “Adapted vocacbularies for generic visual categorization,” *IEEE European Conference on Computer Vision*, 2005.

[10] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, “Unifying discriminative visual codebook generation with classifier training for object category recognition,” *Computer Vision and Pattern Recognition*, 2008.

[11] X.S. Lian, Z. Li, B. Lu, and L. Zhang, “Max-margin dictionary learning for multiclass image categorization,” *European Conference on Computer Vision*, 2010.

[12] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, “Trace ratio vs. ratio trace for dimensionality reduction,” *Computer Vision and Pattern Recognition*, 2007.

[13] Y. Jia, F. Nie, and C. Zhang, “Trace ratio problem revisited,” *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[14] R.O. Duda, P.E. Hart, and D.G. Stork, “Pattern classification, 2nd ed,” 2000, Wiley-Interscience.

[15] A. Iosifidis, A. Tefas, and I. Pitas, “On the optimal class representation in linear discriminant analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1491–1497, 2013.

[16] I. Laptev and T. Lindeberg, “Space-time interest points,” *International Conference on Computer Vision*, pp. 432–439, 2003.

[17] G. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[18] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local svm approach,” *International Conference on Pattern Recognition*, pp. 32–36, 2004.

[19] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” *Computer Vision and Pattern Recognition*, 2009.

[20] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[21] N. Gkalelis, H. KIm, A. Hilton, N. Nikolaidis, and I. Pitas, “The i3dpost multi-view and 3d human action/interaction database,” *Conference on Visual Media Production*, pp. 159–168, 2009.

[22] A. Kovashka and K. Grauman, “Learning a hierarchy of discriminative spacetime neighborhood features for human action recognition,” *Computer Vision and Pattern Recognition*, 2010.

[23] S. Sadanand and J.J. Corso, “Action bank: A high-level representation of activity in video,” *Computer Vision and Pattern Recognition*, 2012.

[24] A. Iosifidis, A. Tefas, and I. Pitas, “Regularized extreme learning machine for multi-view semi-supervised action recognition,” *Neurocomputing*, vol. 145, pp. 250–262, 2014.

[25] J. Yin and Y. Meng, “Human activity recognition in video using a hierarchical probabilistic latent model,” *Computer Vision and Pattern Recognition*, 2010.

[26] A. Klaser, M. Marszalek, and S. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” *British Machine Vision Conference*, pp. 995–1004, 2009.

[27] W. Yang, Y Wang, and Mori. G., “Recognizing human actions from still images with latent poses,” *Computer Vision and Pattern Recognition*, 2010.

[28] M.B. Kaaniche and F. Bremond, “Gesture recognition by learning local motion signatures,” *Computer Vision and Pattern Recognition*, 2010.

[29] M.S. Ryoo and J.K. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” *International Conference on Computer Vision*, 2009.

[30] H. Bilen, V.P. Namboodiri, and L.V. Gool, “Action recognition: a region based approach,” *Applications of Computer Vision*, 2011.

[31] Y. Wang and G. Mori, “Human action recognition by semilatent topic models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762–1774, 2009.

[32] M. Holte, B. Chakraborty, J. Gonzalez, and T. Moeslund, “A local 3d motion descriptor for multi-view human action recognition from 4d spatio-temporal interest points,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 553–565, 2012.