LEARNING A HIERARCHICAL SPATIO-TEMPORAL MODEL FOR HUMAN ACTIVITY RECOGNITION

[1]*Wanru Xu,* [1] *Zhenjiang Miao,* [2] *Xiao-Ping Zhang,* [1] *Yi Tian*

[1]Institute of Information Science, Beijing Jiaotong University, China
[2] Department of Electrical and Computer Engineering, Ryerson University, Canada

ABSTRACT

Recent works have shown that hierarchical models lead to significant improvement in human activity recognition, which can not only enhance descriptive capability, but also improve discriminative power. However, most existing methods exploit just one of the two advantages. In this paper, a new hierarchical spatio-temporal model (HSTM) is proposed to integrate feature learning into two-layer hierarchical classification model simultaneously. On the one hand, the two-layer model has sufficient descriptive capability. The bottom layer aims at capturing spatial relations in each frame and learning high-level representations, and the top layer utilizes these learned features to characterize temporal relations in the whole video sequence. On the other hand, the hierarchical model has strong discriminative power. Both spatial similarity and temporal similarity of activities are measured. Experimental results show that the HSTM can successfully recognize human activities with higher accuracies on one-person actions (KTH and UCF), human-human interactions (CASIA), and human-object interactional activities (Gupta).

Index Terms— Activity recognition, hidden conditional random field, hierarchical structure, spatio-temporal relations

1. INTRODUCTION

For human activity recognition, there are two important issues: activity representation and activity classification. The former is to extract descriptive features to represent activities and the latter is to utilize such type features for corresponding classification. Recent works have shown that hierarchical models can construct long-range and multi-resolution dependencies and lead to significant improvement on both respects. However, most existing methods treat hierarchical models as either hierarchical feature learning [1, 2] or hierarchical classifier [3, 4], so that the advantages of hierarchical models are not fully exploited. In this paper, a new hierarchical spatio-temporal model (HSTM) is proposed to integrate feature learning into hierarchical classification model simultaneously.

For activity representation, some hand-designed features [5, 6, 7, 23] are utilized to capture complex spatio-temporal dynamics in activities, since spatial and temporal dependencies are two key elements in modeling activity. Compared with hand-designed features, hierarchical features learned by deep architectures with multiple hidden layers [8, 9, 10, 11] are more robust and portable. Therefore, we propose a two-layer model where the bottom layer aims at describing spatial relations in each frame and the top layer is utilized to characterize temporal relations in the whole video. Besides, in this two-layer model, some hierarchical features are learned to capture more global or higher level representations. For activity classification, hierarchical classification models [12, 13, 14] can be employed to recognize complex interactional activities due to their strong modeling capability and discriminative power. The limitation is that the relationships between different layers are separated. Rather than making a decision based on the last layer alone, parameters in the HSTM are trained jointly and both two layers have contributions to the final classification.

In summary, the HSTM has three advantages: 1) Integrate hierarchical and structural information into the interpretation process by constructing HSTM on spatial scale and temporal scale, such that all levels of spatio-temporal relations are captured to enhance descriptive capability. 2) Convert raw observations to some high-level semantic representations to combine the flexibility of local features with the discriminability of global features in a consistent multi-layer framework. 3) Derive a joint learning algorithm to train parameters efficiently and effectively, and this makes both spatial similarity and temporal similarity of activities be measured together to obtain superior classification ability. The HSTM fully exerts the advantages of hierarchical models in both activity representation and activity classification.

2. HIERARCHICAL SPATIO-TEMPORAL MODEL

The hierarchical spatio-temporal model is a discriminative model that directly estimates the probability of output con-

This work is supported by the NSFC 61273274, 61672089, 61572064, PXM2016_014219_000025, 973 Program 2011CB302203, National Key Technology R&D Program of China 2012BAH01F03, NSFB4123104 and Engineering Research Council of Canada under Grant RGPIN239031.

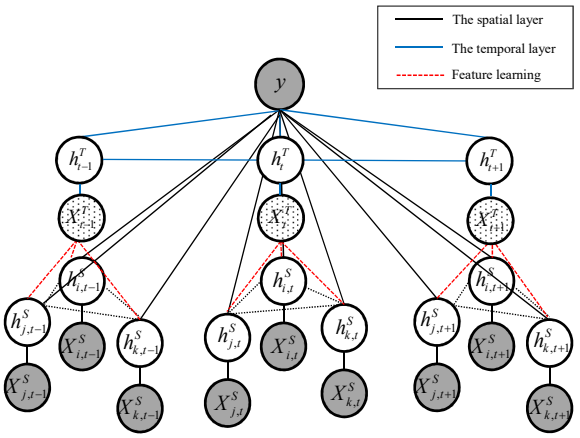

Fig. 1. The graphical representation of HSTM. The HSTM incorporates two layers of hidden nodes denoted as spatial layer and temporal layer, which are connected by black lines and blue lines respectively. In addition, there exists a feature learning process between the two layers, depicted by red dashed lines. In this figure, gray nodes represent observed variables, white nodes denote unobservable variables and shadow nodes are the variables which can be learned. The HSTM models activity in spatial and temporal domain jointly.

ditioned on the whole observations. Our aim is to learn a mapping from observation $\mathbf{X}$ to activity label y, which can match the training data well. This conditional probabilistic model can be formulated as,

$$P(y|\mathbf{X};\boldsymbol{\theta}) = \frac{\sum_{\mathbf{h}\in\mathcal{H}} \exp\left(E(y,\mathbf{h},\mathbf{X};\boldsymbol{\theta})\right)}{\sum_{\mathbf{h}\in\mathcal{H},\hat{y}\in\mathcal{Y}} \exp\left(E(\hat{y},\mathbf{h},\mathbf{X};\boldsymbol{\theta})\right)}, \quad (1)$$

where $\mathcal{Y}$ denotes the possible activity label set and $\hat{y}$ is a member of $\mathcal{Y}$; $\mathcal{H}$ refers to the hidden state set; $\mathbf{h}$ are a set of hidden variables introduced to model complex intra-variations and each hidden variable takes a value from $\mathcal{H}$; and the denominator is a partition function which plays a role of normalization. The $E(y,\mathbf{h},\mathbf{X};\boldsymbol{\theta})$ represents the potential function parameterized by $\boldsymbol{\theta}$, which can model various dependencies among variables depending on the definition form and model structure.

The graphical representation of the new hierarchical spatio-temporal model proposed in this paper is depicted in Fig. 1. It is a two-layer HCRF model, consisting of spatial layer and temporal layer. The spatial relations of local patches within a frame are modeled by a tree structured graph with nodes from the spatial layer. To capture the temporal relations between neighboring frames, hidden nodes in the temporal layer are connected as a chain. Moreover, a feature learning process is adopted to convert raw observations to some high-level semantic representations, which is equivalent to aggregating evidences from local to global level.

2.1. Potential functions of HSTM

To avoid making a decision about which layer is more appropriate for recognition, both two layers have contributions to the final classified result. It can be noted that besides the temporal layer, there also exists a potential function between the spatial layer and the activity label. Therefore, the potential function of HSTM can be defined as the form of a sum,

$$E(y,\mathbf{h},\mathbf{X};\boldsymbol{\theta}) = E^T(y,h^T,X^T;\theta^T) + \sum_{t=1}^{K} E^S(y,h_t^S,X_t^S;\theta^S), \quad (2)$$

where K denotes the number of frames in an activity; X_t^S refer to the observations in the spatial layer at time t and h_t^S are their corresponding spatial hidden variables; X^T denote the observations in the temporal layer and h^T are their corresponding temporal hidden variables; θ^S and θ^T are the parameters in spatial and temporal layers respectively. The definitions of spatial potentials and temporal potentials are given,

$$\begin{aligned} E^S(y,h_t^S,X_t^S;\theta^S) = &\sum_i E_v^S(h_{i,t}^S, X_{i,t}^S;\theta_v^S) + \\ &\sum_{i,j\in\mathcal{E}_t} E_e^S(y,h_{i,t}^S,h_{j,t}^S;\theta_e^S) + \sum_i E_l^S(y,h_{i,t}^S;\theta_l^S), \end{aligned} \quad (3)$$

$$\begin{aligned} E^T(y,h^T,X^T;\theta^T) = &\sum_{t=1}^{K} E_v^T(h_t^T,X_t^T;\theta_v^T) + \\ &\sum_{t=2}^{K} E_e^T(y,h_{t-1}^T,h_t^T;\theta_e^T) + \sum_{t=1}^{K} E_l^T(y,h_t^T;\theta_l^T), \end{aligned} \quad (4)$$

where i, j, t are nodes in HSTM and $\mathcal{E}_t$ denotes the edge set in the spatial layer at time t created by minimum spanning tree.

There are three kinds of potentials in HSTM, which can be defined as a linear production of parameters and features. The feature-related potentials which model the semantic relationships between features and hidden nodes include $E_v^S(h_{i,t}^S,X_{i,t}^S;\theta_v^S) = \sum_n X_{i,t}^S.\mathbb{1}\{h_{i,t}^S = \hbar_n^S\}.\theta_{vn}^S$ and $E_v^T(h_t^T,X_t^T;\theta_v^T) = \sum_n X_t^T.\mathbb{1}\{h_t^T = \hbar_n^T\}.\theta_{vn}^T$. Where $\hbar_n^S$ and $\hbar_n^T$ are the n-th spatial and temporal hidden state respectively; $X_{i,t}^S$ refers to the i-th observation in the spatial layer at time t and $h_{i,t}^S$ is its corresponding hidden node; X_t^T refers to the observation in the temporal layer at time t and h_t^T is its corresponding hidden node; θ_v^S are the parameters of spatial feature-related potential, and $X_{i,t}^S.\theta_{vn}^S$ can be interpreted as how likely the local patch $X_{i,t}^S$ is assigned as the hidden state $\hbar_n^S$; θ_v^T are the parameters of temporal feature-related potential, and $X_t^T.\theta_{vn}^T$ describes how likely the frame X_t^T is assigned as the hidden state $\hbar_n^T$; $\mathbb{1}\{h_{i,t}^S = \hbar_n^S\}$ denotes an indicator function, which is 1 if $h_{i,t}^S = \hbar_n^S$; otherwise it is 0. The activity-related potentials which evaluate the compatibilities between activities and hidden nodes contain $E_l^S(y,h_{i,t}^S;\theta_l^S) = \sum_n\sum_{\hat{y}\in\mathcal{Y}}\mathbb{1}\{y=\hat{y}\}.\mathbb{1}\{h_{i,t}^S = \hbar_n^S\}.\theta_{l\hat{y}n}^S$ and $E_l^T(y,h_t^T;\theta_l^T) = \sum_n\sum_{\hat{y}\in\mathcal{Y}}\mathbb{1}\{y=\hat{y}\}.\mathbb{1}\{h_t^T = \hbar_n^T\}.\theta_{l\hat{y}n}^T$.

Where θ_l^S are the parameters of spatial activity-related potential, and its entry $\theta_{l\hat{y}n}^S$ computes how possible the activity $\hat{y}$ contains a local patch with hidden state $\hbar_n^S$; θ_l^T are the parameters of temporal activity-related potential, and its entry $\theta_{l\hat{y}n}^T$ represents how possible the activity $\hat{y}$ contains a frame with hidden state $\hbar_n^T$. The structure-related potentials model the motion constraints of a pair of hidden states in a specific activity class, consisting of $E_e^S(y, h_{i,t}^S, h_{j,t}^S; \theta_e^S) = \sum_{\hat{y}\in\mathcal{Y}} \sum_n \sum_m \mathbb{1}\{h_{i,t}^S = \hbar_n^S\} \cdot \mathbb{1}\{h_{j,t}^S = \hbar_m^S\} \cdot \mathbb{1}\{y = \hat{y}\} \cdot \theta_{e\hat{y}nm}^S$ and $E_e^T(y, h_{t-1}^T, h_t^T; \theta_e^T) = \sum_n \sum_m \sum_{\hat{y}\in\mathcal{Y}} \mathbb{1}\{h_t^T = \hbar_n^T\} \cdot \mathbb{1}\{h_{t-1}^T = \hbar_m^T\} \cdot \mathbb{1}\{y = \hat{y}\} \cdot \theta_{e\hat{y}nm}^T$. Where θ_e^S are the parameters of spatial structure-related potential, and its entry $\theta_{e\hat{y}nm}^S$ estimates how likely a frame contains a pair of local patches with hidden states $\hbar_n^S$ and $\hbar_m^S$, when given the activity $\hat{y}$; θ_e^T are the parameters of temporal structure-related potential, and its entry $\theta_{e\hat{y}nm}^T$ measures how possible a video contains two consecutive frames with hidden states $\hbar_n^T$ and $\hbar_m^T$ conditioned on the activity $\hat{y}$. Intuitively, the spatial hidden state associates with the "body part label", while the temporal hidden state corresponds to the "pose label".

2.2. Learning two high-level features

In this paper, observations consist of raw features and high-level features. STIPs [15] are extracted as raw features and $X_{i,t}^S$ denotes the feature vector describing appearance of the $i\text{-}th$ local patch at time t. The high-level feature X_t^T learned from the spatial layer is the overall characterization of the $t\text{-}th$ frame. Since the learned spatial hidden variables are compact and semantic, they are used as the basic elements to generate the high-level representations. The X_t^T can be decomposed as the individual features $\Gamma^v(h_t^S) = \{\Gamma_{n,y}^v(h_t^S) | \hbar_n^S \in \mathcal{H}^S, y \in \mathcal{Y}\}$ and the interactional features $\Gamma^e(h_t^S) = \{\Gamma_{n,m,y}^e(h_t^S) | \hbar_n^S, \hbar_m^S \in \mathcal{H}^S, y \in \mathcal{Y}\}$, which characterize high-level components and their spatial dependencies respectively,

$$\begin{aligned}\Gamma_{n,y}^v(h_t^S) &= \sum_i p(h_{i,t}^S = \hbar_n^S | y, X_t^S, \theta^S) \\ &= \frac{\sum_i \sum_{h_{i,t}^S = \hbar_n^S} \exp\left(E^S(y, h_t^S, X_t^S; \theta^S)\right)}{\sum_{h_t^S \in \mathcal{H}^S, \hat{y}\in\mathcal{Y}} \exp\left(E^S(\hat{y}, h_t^S, X_t^S; \theta^S)\right)},\end{aligned} \tag{5}$$

$$\begin{aligned}\Gamma_{n,m,y}^e(h_t^S) &= \sum_{i,j\in\mathcal{E}_t} p(h_{i,t}^S = \hbar_n^S, h_{j,t}^S = \hbar_m^S | y, X_t^S, \theta^S) \\ &= \frac{\sum_{i,j\in\mathcal{E}_t} \sum_{h_{i,t}^S = \hbar_n^S} \sum_{h_{j,t}^S = \hbar_m^S} \exp(E^S(y, h_t^S, X_t^S; \theta^S))}{\sum_{h_t^S \in \mathcal{H}^S, \hat{y}\in\mathcal{Y}} \exp\left(E^S(\hat{y}, h_t^S, X_t^S; \theta^S)\right)},\end{aligned} \tag{6}$$

where $\mathcal{H}^S$ is the spatial hidden state set; the two marginal probabilities $p(h_{i,t}^S = \hbar_n^S | y, X_t^S, \theta^S)$ and $p(h_{i,t}^S = \hbar_n^S, h_{j,t}^S = \hbar_m^S | y, X_t^S, \theta^S)$ can be calculated by belief propagation. From another perspective, it has some similarities to the deep learning paradigm. The spatial layer can be seen as applying a soft-max function over the potentials across all the labels at each local patch and the high-level features can be obtained by a process which is something like structured pooling.

2.3. Training and inference for HSTM

In contrast with other hierarchical models, we jointly optimize model parameters for HSTM. We run Quasi-Newton on the log-likelihood to learn the optimal solution. Given training videos $\{\mathbf{X}_i, y_i\}_{i=1}^M$, the objective function is defined as,

$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^M L(i, \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^M \log \frac{\sum_{\mathbf{h}\in\mathcal{H}} \exp\left(E(y_i, \mathbf{h}, \mathbf{X}_i; \boldsymbol{\theta})\right)}{\sum_{\mathbf{h}\in\mathcal{H}, y\in\mathcal{Y}} \exp\left(E(y, \mathbf{h}, \mathbf{X}_i; \boldsymbol{\theta})\right)} - \frac{1}{2\sigma^2}||\boldsymbol{\theta}||^2.\end{aligned} \tag{7}$$

The first term denotes the log-likelihood on training data and the last is a Gaussian regularization term.

The traditional training algorithm is not tractable for the HSTM due to its hierarchical structure. Therefore, a bottom-to-up strategy is adopted to jointly estimate parameters efficiently and effectively, which can limit the computational complexity to linear to the number of layers. We can reformulate the objective function,

$$\begin{aligned}\max_{\boldsymbol{\theta}} L(i, \boldsymbol{\theta}) &= \max_{\boldsymbol{\theta}} \log \frac{\sum_{\mathbf{h}\in\mathcal{H}} \exp\left(E(y_i, \mathbf{h}, \mathbf{X}_i; \boldsymbol{\theta})\right)}{\sum_{\mathbf{h}\in\mathcal{H}, y\in\mathcal{Y}} \exp\left(E(y, \mathbf{h}, \mathbf{X}_i; \boldsymbol{\theta})\right)} \\ &\geq \max_{\theta^T} \log \frac{\prod_t p_{i,y_i,t}^S \cdot (\sum_{h^T\in\mathcal{H}^T} \exp(E^T(y_i, h^T, X^T; \theta^T)))}{\sum_{y\in\mathcal{Y}} \prod_t p_{i,y,t}^S \cdot (\sum_{h^T\in\mathcal{H}^T} \exp(E^T(y, h^T, X^T; \theta^T)))} \\ &= \max_{\theta^T} \mathcal{L}(i, \boldsymbol{\theta} | \theta^{S^*}),\end{aligned} \tag{8}$$

where only the $i\text{-}th$ training sample $\{\mathbf{X}_i, y_i\}$ is considered for description convenience and the regularization term is also omitted. $\mathcal{L}(i, \boldsymbol{\theta} | \theta^{S^*})$ is an approximation of $L(i, \boldsymbol{\theta})$, when given θ^{S^*}. The well optimized marginal posterior probability of the $t\text{-}th$ frame in the $i\text{-}th$ video conditioned on class y is defined as $p_{i,y,t}^S = \sum_{h_t^S\in\mathcal{H}^S} \exp\left(E^S(y, h_t^S, X_t^S; \theta^{S^*})\right)$ and the parameters in the spatial layer can be estimated by,

$$\theta^{S^*} = \arg\max_{\theta^S} \sum_{t=1}^K \log \frac{\sum_{h_t^S\in\mathcal{H}^S} \exp\left(E^S(y_i, h_t^S, X_t^S; \theta^S)\right)}{\sum_{h_t^S\in\mathcal{H}^S, y\in\mathcal{Y}} \exp\left(E^S(y, h_t^S, X_t^S; \theta^S)\right)}. \tag{9}$$

In fact, the parameters are optimized to maximize a lower bound of likelihood function of the complete HSTM by splitting model into local layers and integrating statistics over all of them. The training procedure is outlined in Algorithm 1.

Table 1. Accuracy of the HSTM compared to other related methods on the KTH, UCF, CASIA and Gupta video dataset.

KTH	Accuracy	UCF	Accuracy	CASIA	Accuracy	Gupta	Accuracy
HCRF [20]	92.51%	CNN [1]	75.8%	CHMM [21]	76.14%	HOI features [22]	93%
M^3HBM [13]	97.99%	ISA [8]	86.5%	CODHMM [19]	83.28%	Bayesian model [18]	93.34%
The HSTM	98%	**The HSTM**	89.33%	**The HSTM**	95.24%	**The HSTM**	96.30%

Algorithm 1 Training Procedure for the HSTM.

1: Estimating the spatial parameters θ^{S^*} as (9);
2: Learning the high-level features for the temporal layer, as (5) and (6);
3: Calculating the marginal posterior probabilities $p^S_{i,y,t}$;
4: Accumulating the approximation of the complete log-likelihood $\mathcal{L}_1(\boldsymbol{\theta}|\theta^{S^*}) = \sum_{i=1}^{M} \mathcal{L}_1(i, \boldsymbol{\theta}|\theta^{S^*})$;
5: Utilizing Quasi-Newton algorithm to estimate model parameters $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}^* \setminus \theta^{S^*}} \mathcal{L}_1(\boldsymbol{\theta}|\theta^{S^*})$;

Given the optimal parameters $\boldsymbol{\theta}^*$, prediction of a new test video $\mathbf{X}$ is to select y that could maximize the conditional probability of our model $y^* = \arg\max_{y \in \mathcal{Y}} P(y|\mathbf{X}; \boldsymbol{\theta})$.

3. EXPERIMENTS

Our proposed model is a unified framework to recognize one-person actions and interactional activities. We evaluate performance of the HSTM on three tasks with four standard benchmark datasets: one-person action recognition (KTH [16] and UCF [17]), human-human interactional activity recognition (CASIA [19]) and human-object interactional activity recognition (Gupta [18]). Except for the CASIA, where five trajectory based features [19] are extracted, we adopt the same raw features (STIPs) for the other three datasets.

In order to comprehensively evaluate performance of the HSTM, we compare it with several related models on the four benchmark datasets shown in Table 1. These comparable approaches can be divided into three categories: 1) The related graphical models: HCRF based models (eg.[20]) and HMM based models (eg.[19, 21]). 2) The hierarchical feature learning: unsupervised hierarchical feature learning (eg.[8]) and deep learning model (eg.[1]). 3) The hierarchical classification model: multi-feature max-margin hierarchical Bayesian model (eg.[13]). It can be seen that the HSTM is comparable to all the state-of-the-art methods on the four datasets. In particular, our model achieves a near-perfect accuracy on the KTH (98%) and a huge improvement on the CASIA (about 12%). It strongly demonstrates that our HSTM combining the advantages of both hierarchical feature learning and hierarchical classification has stronger discriminative power and descriptive capability.

Table 2. Compare the HSTM with its one-layer sub-models.

Algorithm	KTH	UCF	CASIA	Gupta
PFSL	34.35%	17.71%	58.98%	24.07%
PVSL	37.17%	18%	84.52 %	28.92%
RFTL	76.5%	58.67%	51.19%	59.26%
LFTL	97.67%	86.67%	92.86 %	88.89%
The HSTM	98%	89.33%	95.24 %	96.30%

For more detailed analysis, we evaluate whether our HSTM is indeed better than one-layer models. To this end, we compare the HSTM with the single spatial layer (PFSL and PVSL) and the single temporal layer (RFTL and LFTL), where Per-Frame in Spatial Layer is to classify every frame independently; Per-Video in Spatial Layer is to achieve the whole video label by majority voting; Raw Features in Temporal Layer is to use the raw features as the observations of the temporal layer; Learned Features in Temporal Layer is the method which only applies the learned high-level features and ignores the frame-level marginal posterior probabilities. From the comparisons with the HSTM and its one-layer sub-models in Table 2, it can find that the HSTM achieves the highest accuracy and there is no one-layer representation that is as discriminative as the hierarchical representation.

4. CONCLUSIONS

This paper proposes a new hierarchical spatio-temporal model for both one-person action recognition and interactional activity recognition by modeling spatial constraints and temporal constraints simultaneously. In the HSTM, the advantages of hierarchical models on the aspects of both activity representation and activity classification are fully exploited. Firstly, the descriptive capability is enhanced not only by integrating all levels of spatio-temporal relations but also by combining the flexibility of local features with the discriminability of global features. Secondly, a joint learning algorithm with bottom-to-up strategy is derived to train parameters efficiently and effectively. Both spatial similarity and temporal similarity of activities are measured together to obtain superior classification ability. We evaluate the HSTM on the KTH, UCF, CASIA and Gupta video dataset and obtain better recognition results compared with state-of-the-art methods.

5. REFERENCES

[1] G. Gkioxari and J. Malik, "Finding action tubes", in *Computer Vision and Pattern Recognition*. IEEE, 2015.

[2] L. Wang, T. Liu and G. Wang, "Video tracking using learned hierarchical features", *IEEE Transactions on Image Processing*, vol.24, no.4, pp.1424–1435, 2015.

[3] Q. Huang, M. Han and B. Wu, "A hierarchical conditional random field model for labeling and segmenting images of street scenes", in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp.1953–1960.

[4] L. Zhang, Z. Zeng and Q. Ji, "Probabilistic image modeling with an extended chain graph for human activity recognition and image segmentation", *IEEE Transactions on Image Processing*, vol.20, no.9, pp.2401–2413, 2011.

[5] M. Bregonzio, S. Gong and T. Xiang, "Recognising Action as Clouds of Space-Time Interest Points", in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp.1948–1955.

[6] S. Song, N.M. Cheung and V. Chandrasekhar, "Egocentric activity recognition with multimodal fisher vector", in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2016, pp.2717–2721.

[7] J. Wang, Z. Chen and Y. Wu, "Action recognition with multiscale spatio-temporal contexts", in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp.3185–3192.

[8] Q.V. Le, W.Y. Zou and S.Y. Yeung, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis", in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp.3361–3368.

[9] Y. Song, L.P. Morency and R. Davis, "Action recognition by hierarchical sequence summarization", in *Computer Vision and Pattern Recognition*. IEEE, 2013, pp.3562–3569.

[10] L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors", in *Computer Vision and Pattern Recognition*. IEEE, 2015.

[11] Y. Liu, L. Qin and Z. Cheng, "DA-CCD: A novel action representation by Deep Architecture of local depth feature", in *International Conference on Image Processing*. IEEE, 2014, pp.833–837.

[12] X. Wang, and Q. Ji, "Video Event Recognition with Deep Hierarchical Context Model", in *Computer Vision and Pattern Recognition*. IEEE, 2015, pp.4418–4427.

[13] S. Yang, C. Yuan and B. Wu, "Multi-Feature Max-Margin Hierarchical Bayesian Model for Action Recognition", in *Computer Vision and Pattern Recognition*. IEEE, 2015, pp.1610–1618.

[14] T. T. Thanh, F. Chen and K. Kotani, "Automatic extraction of semantic features for real-time action recognition using depth architecture networks", in *International Conference on Image Processing*. IEEE, 2014, pp.1540–1544.

[15] P. Dollr, V. Rabaud and G. Cottrell, "Behavior recognition via sparse spatio-temporal features", in *Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp.65–72.

[16] C. Schuldt, I. Laptev and B. Caputo, " Recognizing human actions: a local SVM approach", in *International Conference on Pattern Recognition*. IEEE, 2004, vol. III, pp.32–36.

[17] M.D. Rodriguez, J. Ahmed and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition", in *Computer Vision and Pattern Recognition*. IEEE, 2008, pp.1–8.

[18] A. Gupta, A. Kembhavi and L.S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.31, no.10, pp.1775–1789, 2009.

[19] P. Guo, Z. Miao and X.P. Zhang, "Coupled observation decomposed hidden markov model for multiperson activity recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol.22, no.9, pp.1306–1320, 2012.

[20] Y. Wang and G. Mori, "Learning a discriminative hidden part model for human action recognition", in *Advances in Neural Information Processing Systems*. pp.1721–1728, 2009.

[21] M. Brand, N. Oliver and A. Pentland, "Coupled hidden Markov models for complex action recognition", in *Computer Vision and Pattern Recognition*. IEEE, 1997, pp.994–999.

[22] A. Prest, V. Ferrari and C. Schmid, "Explicit modeling of human-object interactions in realistic videos", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.35, no.4, pp.835–848, 2013.

[23] J. Sun, X. Wu and S. Yan, "Hierarchical spatio-temporal context modeling for action recognition", in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp.2004–2011.