DSPM: DYNAMIC STRUCTURE PRESERVING MAP FOR ACTION RECOGNITION

Anonymous ICME submission

ABSTRACT

In this paper, a Dynamic Structure Preserving Map (DSPM) is proposed to effectively recognize human actions in video sequences. Inspired by the latest feature learning methods, we modified and improved the adaptive learning procedure in self-organizing map (SOM) to capture dynamics of best matching neurons through Markov random walk. The DSPM can learn implicit spatial-temporal correlations from sequential action feature sets and preserve the intrinsic topologies characterized by different human motions. A further advantage of DSPM is its ability to learn low-level features in challenging video data. The projection from high dimensional action features to low dimensional latent neural distribution significantly reduces the computational cost and data redundancy in the recognition process. The effectiveness and robustness of the proposed method is verified through extensive experiments on several benchmark datasets.

Index Terms— Human action recognition; spatio-temporal dependency; self-organizing map; Markov random walk

1. INTRODUCTION

Human action recognition has attracted much attention in the fields of computer vision and machine learning in recent years [1]. Many previous works have focused on augmenting the feature descriptions, such as proposing stronger feature sets and combining different features [4], or improving action recognition models, such as clustering and classification for scene analysis or abnormal events detection [2]. The analysis of human actions in a video sequence is challenging, because the recognition system is required to extract implicit properties including spatio-temporal coherence, behavior dynamics, and shape deformation. The action feature extraction from a video sequence is different from static image analysis, since spatio-temporal variation might result in meaningful behavior patterns. For example, the changes in human motion orientation or gesture during a specified time interval may indicate what actions might have occurred. In real world applications, irregular behaviors or environments should also be taken into account, which requires the dynamic model to adapt to unexpected factors.

Recently, unsupervised feature learning methods [3] have shown promising potentials for human action recognition. In this paper, we propose a Dynamic Structure Preserving Map (DSPM) to integrate automatic feature learning with spatial-temporal modeling for human action recognition. The unique properties of our proposed method can be summarized as follows:

First, DSPM is able to learn low-level features and produce a generative model to represent the dynamic topological structure. Instead of extracting carefully selected features, our method can automatically learn intrinsic characteristics from raw optical flow field for action recognition. Extending to the self-organizing map (SOM) model [5], DSPM accumulates dynamic behavior of best-matching units (BMUs) to adjust their synaptic neuron weights, which can effectively capture the temporal information.

Second, DSPM can aggregate the spatio-temporal clustering while simultaneously preserve underlying topological structure. Characterized by the parameters of latent neural distribution and neighborhood kernel function, the highly relevant spatio-temporal correlations for each action feature set are adaptively preserved in a 2-D lattice of neurons.

Third, DSPM provides an effective way to reduce the dimensionality of input raw feature set, such as dense optical flow, to represent human motions in videos. Through the non-linear mapping procedure, DSPM can reduce the computational cost and data redundancy in action recognition.

The remainder of this paper is organized as follows. Related works of human action recognition are introduced in Section 2. In Section 3, the mechanisms of related learning methods are analyzed. We introduce our DSPM method and present the detailed learning procedures in Section 4. In Section 5, experimental results are presented and discussed. Finally, a conclusion is provided in the Section 6.

2. RELATED WORK

As a typical classification problem, feature extraction plays an essential role in the action recognition. Due to the intrinsic sequential property, many spatio-temporal features, such as STV [6], STIP [7, 8], HOSVD [9] have been developed. Besides the spatio-temporal property, feature sets with multiple hierarchies are also introduced for action recognition. Sun $et\ al.$ [10] modeled spatio-temporal context information in a hierarchical structure. Three levels of context were established in ascending order of abstraction: point-level context, intra-trajectory context, and inter-trajectory context. Gilbert

et al. [11] introduced a novel approach to use very dense corner features, which were spatially and temporally grouped in a hierarchical process to produce an overcomplete compound feature set. In addition, the spatio-temporal feature set is also combined with other features, such as shapes [12], to make the action more descriptive.

Besides augmenting the features, different machine learning algorithms also have been introduced to improve the human action recognition performance. Zhu *et al*. [13] adopted multi-class support vector machine (SVM) with linear kernels. Schuldt *et al*. [14] used local space-time features for recognizing complex motion patterns. They constructed video representations in terms of local space-time features and integrated these representations with SVM classification schemes for action recognition. To improve the robustness, a Multiple Kernel Learning with Augmented Features (AFMKL) was proposed to learn an adapted classifier based on multiple kernels and pre-learned classifiers of other action classes in [15]. Fathi *et al*. [16] classified the input video sequence into one of the discrete action classes. The low-level motion features were used as the weak classifiers. The mid-level shape features were constructed from low-level gradient features using AdaBoost. To aggregate the information from different parts of the video sequence, AdaBoost was used for a second time to train the final classifier from the mid-level motion features.

Rather than computing the hand-engineered features or introducing complex classification models, we adopted the feature learning concept in [4] and [17], together with SOM model, to build DSPM to persevere the underlying highly relevant structure both in spatial and temporal dimension.

3. DYNAMIC STRUCTURE PRESERVING MAP FOR ACTION RECOGNITION

It is a complex process to analyze the correlation and variation across space and time. There are limitations on the estimation of traditional state-space models, since the high dimensional parameters may lead to complex dependency structures. Based on the clusters on the spatio-temporal feature map defined in DSPM, the parameters of the latent space model are estimated. The ensemble learning based on EM further enhances the dynamic model to yield better performance. The classifier with highest likelihood will be selected to predict class label. The training procedure of DSPM can be illustrated in Fig. 1.

3.1. Self-organizing Map

SOM [5] is considered as a powerful neural network model in unsupervised learning, which can extract certain implicit knowledge without human intervention or empirical evidence. Given the input data sequence $X = \{x_1, ..., x_n\}$ and synaptic neuron weight m_j, $j \in \{1, ..., N_s\}$, N_s is the total number of the neurons on the map. The procedure of searching the best-matching unit (BMU) can be expressed as (1).

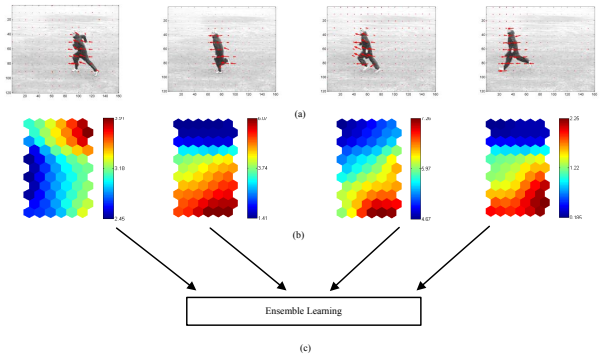

Fig. 1: Training procedure of the proposed dynamic model. (a) Optical flow is extracted from each action video sequences. Given two consecutive frames, optical flow is computed at each pixel, and sampled with a 10×10 grid. For instance, the frame size of KTH data set is 160×120, after optical flow computing, the size of optical flow field for each frame is 16×12×2. The third dimension 2 indicates the magnitude and direction of optical flow. (b) Example DSPMs describing spatio-temporal patterns. The colors of grid represent the distances of various motions on DSPM. (c) The EM based ensemble learning is adopted to predict the action class.

$$b_i = \arg\min_j \parallel x_i - m_j \parallel \tag{1}$$

Gaussian neighborhood kernel function defined in (2) is used to constrain the neighborhood scope of the BMU.

$$h_{j,b_i}(t) = \exp(-\frac{d^2_{j,b_i}}{2\sigma^2(t)}) \tag{2}$$

where $d_{j,b_i} = \parallel r_j - r_{b_i} \parallel$, $\sigma(t) = \sigma_0(\frac{\sigma_1}{\sigma_0})^{\frac{t}{N_c}}$, r_j is a 2-D position vector of neuron j; t represents the training time; N_c denotes the convergence iterations; σ_0 and σ_1 are initial and terminal neighborhood radius, respectively.

An adaptive learning rule updates the synaptic neuron weight $m_{j,t+1}$ according to (3).

$$m_{j,t+1} = m_{j,t} + \alpha(t)h_{j,b_i}(t)(x_i - m_{j,t}) \tag{3}$$

where $\alpha(t) = \alpha_0(\frac{\alpha_1}{\alpha_0})^{\frac{t}{N_c}}$, α_0 and α_1 represent the initial and terminal learning rate, respectively.

3.2. Dynamic Structure Preserving Map

In SOM, the neighborhood function can be only used to preserve the spatial topology. Several extensions to SOM, including Temporal Kohonen map (TKM) and recurrent self-organizing map (RSOM) [18], have been proposed to adaptively model a data distribution over time on non-stationary input sequences. Although TKM preserves a trace of the past activation in terms of weighted sum, the weights are only updated towards the last frame sample of the input sequence based on the convention SOM update rule. RSOM provides a

consistent update rule for the network parameters. The main objective of TKM and RSOM is to follow the trend of the temporal sequence while smoothing out temporary volatilities. These method emphasize more on the latest samples, and eventually remove the influence from old samples. On the contrary, DSPM intends to capture the whole dynamic patterns within the data sequence. We improve the learning rule of DSPM based on (3). The input sequential samples, after some simple cleaning operation, have the same importance and contribute evenly to the model from the beginning to the end. The resulting DSPM with the complete spatio-temporal information is then used in classification. In particular, the neuron transition probabilities in DSPM can describe the temporal dynamics from the training video sequences. DSPM models sequential dynamics by introducing Markov process to capture neuron transition probabilities between every two time samples. It is similar to Markov random walk [19] on graph, where at each step the walk jumps from one place to another based on specified probability distribution. The parameters of Markov process are used in neuron update and model classification.

Suppose the video sequences $X = \{X_1, ..., X_S\}$, the i_{th} sequence $X_i = \{x_{i,1}, ..., x_{i,t}, x_{i,t+1}, ..., x_{i,T}\}$, where $x_{i,t}$ is the video frame data at time t. In DSPM, we have N_s neurons on the lattice map. $p_{i,j}$ is the transition probability from BMU i at the time t to BMU j at the time $t+1$.

$$p_{i,j} = \frac{K(i,j)}{\sum_{m \in N_s} K(i,m)} \tag{4}$$

The kernel function K is: $K(i,j) = exp(-d(i,j)/\alpha)$, where $d(i,j)$ is Manhattan distance between BMU i at the time t to BMU j at the time $t+1$ on the lattice map, α is a constant.

Fig. 2 illustrates the adaptive learning rule of DSPM. $m^*_{j,t+1}$ and $m_{j,t}$ are used to update the synaptic weights. We can see that $m^*_{j,t+1}$ can be calculated by $m_{j,t}$ and $m_{j,t+1}$. $m_{j,t}$ means the neuron weight at the previous time. The transition probabilities constrain the variations of neuron weights, which keeps temporal dependencies between $m_{j,t}$ and $m_{j,t+1}$. This formulation means the elastic characteristics of DSPM have effects on both spatial domain as well as temporal domain. The temporal properties depend on neighborhood topology and dynamic information. The synaptic neuron weights can be updated according to (5).

$$m_{j,t+1} = m_{j,t} + \alpha(t) h_{j,b_{i,t}}(t)(x_{i,t} - p_{b_{i,t},b_{i,t+1}} m^*_{j,t+1} - (1 - p_{b_{i,t},b_{i,t+1}}) m_{j,t}) \tag{5}$$

where $m^*_{j,t+1} = (1 - p_{b_{i,t},b_{i,t+1}}) m_{j,t} + p_{b_{i,t},b_{i,t+1}} m_{j,t+1}$, $m^*_{j,t+1}$ can help the target neuron to adaptively learn temporal knowledge from Markov model. The transition probability can teach the neuron how to preserve the temporal information by updating the neuron weight in DSPM.

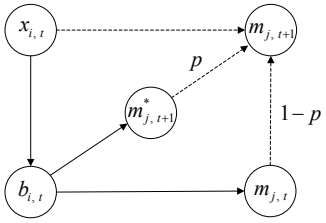

Fig. 2: Adaptive learning rule of DSPM.

Algorithm 1 Spatio-temporal DSPM

Input:
(1) Video feature sequences: $X = \{X_1, ..., X_S\}$, where X_i is a sequence vector $\{x_{i,1}, ..., x_{i,T}\}$
(2) Initial neuron weights: $m_j(t_0)$
$X \leftarrow \frac{X - \min(X)}{\max(X) - \min(X)}$, $X \in [0,1]$
for $i = 1$ to S **do**
  for $t = 1$ to T **do**
    Search BMU $b_{i,t}$ as (1)
    Calculate $p_{b_{i,t},b_{i,t+1}}$ as (4)
    Update $m_{j,t+1}$ as (5)
    $d_{i,t} \leftarrow b_{i,t}$
  end for
end for
Output: Discrete sequences $D^*_i = \{d_{i,1}, ..., d_{i,T}\}$

We take the frame samples of "*bend*" action from 9 persons in the Weizmann dataset in Fig. 3. DSPM can extract the key feature information by spatio-temporal knowledge and statistically measure the dependency by Markov transition probability. The green color grid is the output of DSPM, which can aggregate the key features into the clustering. The x coordinate represents temporal feature in frame number and the y coordinate means the cost on distance between the input video frame and its best matching neuron in DSPM. The red marked circles represent the corresponding cluster in the DSPM. We can see that key features with sparse distribution have a high cost on distance.

4. DYNAMIC MODEL FOR ACTION RECOGNITION

The proposed dynamic model can optimize the parameters and train the ensemble learning model for classification.

We assume the input data $X_t = \{X(x_i; t)\}$, $i = 1, ..., S$, where S is the number of spatial data attributes at the time t. The covariance matrix of The zero-mean Gaussian noise is Δ_t. Θ_t describes the state transition over the time t.

We collect the dynamic model parameters as $\Phi = \{\Theta_t, \Delta_t\}$. The primary goal of this model is to estimate the modeling parameters through expectation-maximization (EM). From Algorithm 1, we can obtain the discrete sequences D^*. The likelihood of the input data sequences can be estimated as (6).

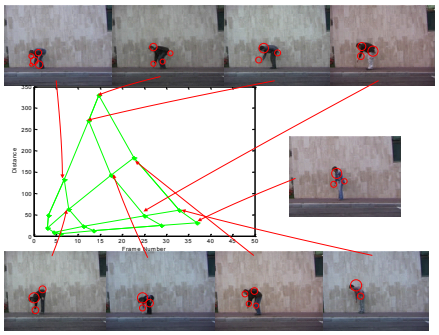

Fig. 3: Spatio-temporal dependency analysis on the key region.

$$P(D^*|\Phi) = \prod_{i=1}^{n} P(D_i^*|\Phi) \tag{6}$$

We can predict the class label based on (7).

$$y = \arg\max_{s_i \in S} \sum_{j=1}^{n} P(D^*|\overline{\Phi_j}, s_j) P(s_j|\overline{\Phi_j}) P(\overline{\Phi_j}) \tag{7}$$

where $\overline{\Phi_j}$ represents one of the alternative models, S is the set of all class labels.

5. EXPERIMENTS

KTH [14], Weizmann action [20] and UCF sport datasets [21] are used to evaluate the performance of the proposed method. To analyze the effects of periodic and non-periodic actions, we calculate optical flow in feature extraction [22]. Optical flow approximates local image motion based on local derivatives in a video sequence, and it can essentially reflect the spatio-temporal variability between two consecutive frames.

5.1. Performance

The performance of the proposed approach can be analyzed through the confusion matrix. In KTH dataset, "$walk$" can be easily recognized with the rate of 98% in Fig. 4, but it is confused by "run" with 2%. "jog" and "run" are both affected by "$walk$". "$handwave$" and "$handclap$" affect the recognition results with each other. From Fig. 5, we can see that our method achieves 100% accuracy for recognizing the actions including "$jack$", "$jump$", "$pjump$", "$side$" and "$walk$". There are some errors in other actions. For example, both "$bend$" and "$wave2$" are the actions with two-hand up in some specified scenarios. The spatial similarities over time make it difficult to achieve high accuracy. As shown in Fig. 6, the sport action recognition is also a challenging task. We can recognize the action "$dive$" with high accuracy, but it becomes more difficult to recognize other actions, such as "run". Although the spatio-temporal dynamic topological structure improves dynamic model to make accurate decision, the false recognition occurs when training frame snapshots or sequences shares the similar variations of spatio-temporal features.

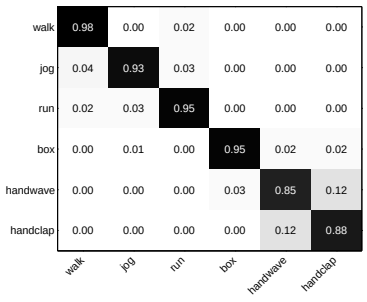

Fig. 4: Confusion matrix on KTH action dataset.

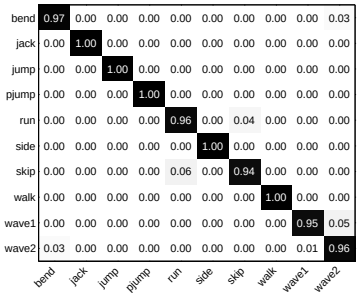

Fig. 5: Confusion matrix on Weizmann action dataset.

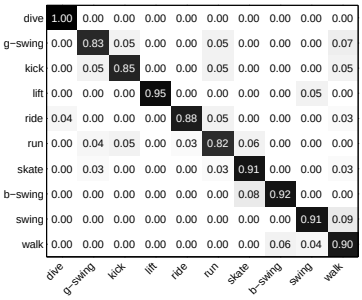

Fig. 6: Confusion matrix on UCF action dataset.

To verify the recognition capability of the proposal method, Table 1 shows the recognition results of many comparable approaches based on KTH, Weizmann and UCF dataset, respectively. On KTH dataset, Wu *et al.* [15] and Kovashka *et al.* [17] achieved the best performance with 94.5%.

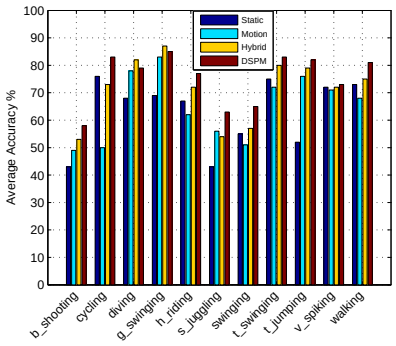

Fig. 7: Recognition performance on YouTube dataset.

Table 1: AVARAGE ACCURACY ON KTH, WEIZMANN, UCF, AND YOUTUBE DATASETS

Method	KTH	Weizmann	UCF	YouTube
Fathi *et al.* [16]	90.5%	100%	-	-
Dollar *et al.* [23]	81.2%	86.7%	-	-
Niebles *et al.* [24]	81.5%	90.0%	-	-
Zhang *et al.* [25]	91.3%	92.9%	-	-
Blank *et al.* [20]	-	100%	-	-
JHuang *et al.* [26]	91.7%	98.8%	-	-
Schuldt *et al.* [14]	71.7%	-	-	-
Laptev *et al.* [7]	91.8%	-	-	-
Klaser *et al.* [27]	91.4%	84.3%	-	-
Campos *et al.* [28]	91.5%	96.7%	80.0%	-
Wang *et al.* [29]	89.0%	97.8%	83.3%	-
Wu *et al.* [15]	94.5%	-	91.3%	-
Kovashka *et al.* [17]	94.5%	-	87.3%	-
Liu *et al.* [30]	93.8%	-	86.5%	71.2%
Le *et al.* [3]	93.9%	-	86.5%	75.8%
Our method	94.2%	98.7%	91.6%	76.5%

Our method can achieve 94.2% on average. On Weizmann dataset, Blank [20] and Fathi [16] achieved 100%, Jhuang [26] achieved 98.8%, and our method achieved 98.7%. On the most challenging UCF dataset, Kovashka *et al.* [17] and Wu *et al.* [15] achieved 87.3% and 91.3%, respectively. Our method with 91.6% performs better than these methods. The performance of our method is comparable with these state of the art methods on action datasets. Particularly, for more complex dataset, such as UCF sport dataset, our method can effectively improve the recognition performance. But more importantly our method can adaptively learn from low level features, such as optical flow, rather than using strong features. This improves model robustness, and requires less human intervention.

To further analyze the robustness of our proposed method on challenging realistic actions, we compare our method with the work by Liu *et al.* [30] based on UCF YouTube dataset with 11 action categories. This video dataset is very challenging, including mixture of steady and shaky cameras, diversity of background, different viewpoint, various illumination and low resolution. Fig. 7 shows the recognition accuracies of three variations of method in [30] and our DSPM. From Fig. 7, we can see that DSPM outperforms the average recognition accuracy in [30], especially for some difficult scenarios such as "$s\_juggling$", "$swinging$" and "$b\_shooting$". The experiment results indicate that our method performs particularly well in recognizing cyclic actions such as "$s\_juggling$", "$swinging$", "$cycling$", "$t\_jumping$" and "$walking$". The performance is not as good as the hybrid method in [30] on "$diving$", since there is significant data redundancy at the beginning of the sequence. The hybrid method adopts a heuristic pruning strategy to reduce the redundant frames. DSPM has the capability to handle this problem if a simple redundancy detection is employed. The similar performance on "$v\_spiking$" indicates the effectiveness of DSPM to recognize group action of multiple people. [30] are mainly focused on three kind of video feature styles. Through the comparison in table I, DSPM can perform as a competitive method compared with the existing methods.

6. CONCLUSION

In this paper, we proposed a new DSPM as an effective spatio-temporal model to recognize human actions from video sequences. Through learning on low level features, DSPM automatically extracts intrinsic spatio-temporal patterns from the video sequence. DSPM improves the adaptive learning rule with a Markov model on the dynamic behavior of BMUs, which helps to preserve spatio-temporal dynamic topological structure. Through the non-linear mapping, DSPM can reduce computational cost and data redundancy for action recognition. The ensemble learning based on EM is adopted to estimate the latent parameters. In the future work, we will continue to improve DSPM to efficiently recognize more complex human actions from the real-world video datasets.

7. REFERENCES

[1] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[2] T. Duong, H. Bui, D. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-markov model," *CVPR*, pp. 838–845, 2005.

[3] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," *CVPR*, 2011.

[4] H. Lee, R. Grosse, R. Ranganath, and A. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," *ICML*, 2009.

[5] T. Kohonen, "Self-organizing maps," *Springer*, 2001.

[6] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," *CVPR*, 2005.

[7] I. Laptev, "On space-time interest points," *Intl. Journal of Computer Vision*, vol. 64, pp. 107–123, 2005.

[8] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *ECCV*, 2008.

[9] Y. Lui, J. Beveridge, and M. Kirby, "Action classification on product manifolds," *CVPR*, 2010.

[10] J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, and J. Li, "Hierarchical spatio-temporal context modeling for action recognition," *CVPR*, 2009.

[11] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," *ICCV*, 2009.

[12] Y. Ke, R. Sukthankar, and M. Hebert, "Spatio-temporal shape and flow correlation for action recognition," *CVPR*, 2007.

[13] Y. Zhu, X. Zhao, Y. Fu, and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," *ACCV*, pp. 660–671, 2010.

[14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," *ICPR*, 2004.

[15] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," *CVPR*, 2011.

[16] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *CVPR*, 2008.

[17] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," *CVPR*, 2010.

[18] M. Varsta, J. Heikkonen, J. Lampinen, and J. Millan, "Temporal kohonen map and recurrent self-organizing map: analytical and experimental comparison," *Neural Processing Letters*, vol. 13, pp. 237–251, 2001.

[19] M. Szummer and T. Jaakkola, "Partially labeled classification with markov random walks," *NIPS*, 2001.

[20] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *CVPR*, 2005.

[21] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach: a spatio-temporal maximum average correlation height filter for action recognition," *CVPR*, 2008.

[22] D. Fleet and Y. Weiss, "Optical flow estimation," *Handbook of Mathematical Models in Computer Vision, Springer*, pp. 239–258, 2005.

[23] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," *IEEE Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[24] J. Niebles, H. Wang, and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Intl. Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[25] Z. Zhang, Y. Hu, S. Chan, and L. Chia, "Motion context: Anew representation for human action recognition," *ECCV*, pp. 817–829, 2008.

[26] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," *ICCV*, 2007.

[27] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," *BMVC*, 2008.

[28] T. Campos, M. Barnard, K. Mikolajczyk, J. Kittler, F. Yan, W. Christmas, and D. Windridge, "An evaluation of bags-of-words and spatio-temporal shapes for action recognition," *IEEE Workshop Applications of Computer Vision*, pp. 344–351, 2011.

[29] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," *BMVC*, 2009.

[30] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," *CVPR*, 2009.