GENRE CATEGORIZATION OF AMATEUR SPORTS VIDEOS IN THE WILD

Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa

Michigan State University, East Lansing, MI, USA

ABSTRACT

Various sports video genre categorization methods are proposed recently, mainly focusing on professional sports videos captured for TV broadcasting. This paper aims to categorize sports videos in the wild, captured using mobile phones by people watching a game or practicing a sport. Thus, no assumption is made about video production practices or existence of field lining and equipment. Motivated by distinctiveness of motions in sports activities, we propose a novel motion trajectory descriptor to effectively and efficiently represent a video. Furthermore, temporal analysis of local descriptors is proposed to integrate the categorization decision over time. Experiments on a newly collected dataset of amateur sports videos in the wild demonstrate that our trajectory descriptor is superior for sports videos categorization and temporal analysis improves the categorization accuracy further.

Index Terms— Genre categorization, Activity recognition, Trajectory, Amateur sports video, Temporal analysis.

1. INTRODUCTION

Content-based video categorization has a crucial role in making the ever-increasing amount of digital content accessible. Automatic indexing and genre categorization of sports videos, as a large portion of digital contents, is of significance and enables domain-specific analysis of sports videos [1].

Sports video categorization is a challenging problem. There might be large inter-class similarities due to similarity of movements, co-existence of players and audiences, and commonality between playing fields. In addition, intra-class variations, such as different movements within a single sport video, camera angle variations, and distinct speeds of actions by different people, render the categorization task difficult. While most previous works assume that sports videos are captured for TV broadcasting, and thus, happen in specific sports fields [1–3], this work aims to analyze sports videos *in the wild*, i.e., amateur videos captured by mobile phones (Fig. 1). These videos have additional challenges due to the field variations, camera motion, and the unskillful capturing.

Many sports activities have a very distinctive set of motions that can be useful to characterize the sport. Our approach is built upon dense trajectories [4] extracted from the optical flow based tracking. However, the simple trajectory descriptor in [4] does not explicitly encode the shape or temporal dynamics of trajectories. It is important to have an efficient and effective trajectory shape descriptor that is robust to camera angle variations and different speeds of actions. Based on these requirements, we propose a novel Orientation based Camera angle Invariant Trajectory descriptor called OCIT, which is both compact and discriminative.

Fig. 1. Categorization of sport videos captured by mobile phones.

Furthermore, if the temporal ordering or dynamic of the trajectories is not represented, some sports may be confused with each other. Hence, it is beneficial to analyze the trajectories in a temporal framework. In this paper, a temporal analysis (TA) method is proposed to capture local descriptors in overlapping blocks over time and fuse the analysis results from all blocks for making the final categorization decision.

We collect a large dataset of amateur videos captured by users of a leading sports video mobile app. Experiment results on this dataset show that OCIT outperforms displacement-based trajectory descriptor used in [4,5]. Also, TA of different types of descriptors improves the performances of the Bag of Words (BoW) [6] method using the same descriptors.

In summary, this paper makes three contributions: (i) To the best of our knowledge, this is the first work on categorizing sports videos in the wild; (ii) A novel trajectory descriptor is proposed to capture trajectory information discriminatively and efficiently; (iii) A temporal analysis approach is presented to integrate the categorization of local descriptors over time.

Previous Work Most prior works assume that sports occur in sports arena (thus the existence of specific equipment and field lining), and videos are captured by professional TV broadcast crew. Duan et al. classify TV broadcasting sports videos via motion, color, and shot length features [2]. In [1], the dominant motion and color in each frame is used to classify 3 sports genres. In [7], a hierarchical SVM is used to categorize sports genres by employing temporal and spatial features. Assuming distinct playing field for different sports, histograms of edge direction and intensity are used to categorize 5 sports genres in [8]. In [9] SIFT features of the sample

U.S. Government work not protected by U.S. copyright

frames and BoW [6] are used to categorize sports videos.

Sports videos in the wild are more challenging for visual analysis than broadcasting videos. Firstly, the static image context is less discriminative. Secondly, the camera angle variation is enormous and videos are affected by camera motion. Thirdly, multiple activities may appear in a single video. Finally, cluttered backgrounds increase analysis difficulty. Given these challenges, for sports videos in the wild, we prefer motion-based analysis to context-based analysis.

As a relevant topic, activity/action recognition differs in nature to sports video categorization. While the former aims to recognize specific actions *separately*, in the latter a wide range of activities fall into a *single* genre, leading to a greater intra-class variance. Also, unlike our dataset, action recognition datasets are usually comprised of short videos that precisely encapsulate the action of interest. Activity recognition works can be categorized in recognition from still images [10–12] and videos [13]. They can also be divided to context [9, 12] or motion based methods [4, 14–16]. In the latter, either space-time features [14,15] or trajectories of motions are extracted [4, 16–18]. For both, the dense sampling outperforms interest-based sampling [4, 19]. Our work is a new development along the trajectory-based method that by introducing a novel trajectory descriptor and temporal analysis, improves genre categorization of sports videos in the wild.

2. OUR PROPOSED APPROACH

In our method, motion is analyzed by extracting dense trajectories [4]. To robustly analyze videos in the presence of camera motion, frame by frame motion stabilization is first achieved by matching interest points on consecutive frames and applying the RANSAC algorithm [20] to obtain the affine transformation between consecutive frames.

2.1. Dense trajectory and descriptors

As proposed in [4], dense trajectories are extracted at multiple spatial scales. Each point $p_t = (x_t, y_t)$ at frame t is tracked to the next frame $t+1$ by performing median filtering in a dense optical flow field $\mathbf{W} = (u_t, v_t)$,

$$p_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (K * \mathbf{W})|_{(\overline{x_t}, \overline{y_t})}, \quad (1)$$

where K is the median filtering kernel and $(\overline{x_t}, \overline{y_t})$ is the rounded position of (x_t, y_t). Trajectories are started from the sample points on a grid spaced by W pixels. The length of each trajectory is limited to L, and after reaching this length, the trajectory is removed from the tracking process and new sample points are tracked.

Displacement-based trajectory descriptor Shape of the trajectories can be used as a representative feature. In [4], the displacements of trajectory, $\Delta p_t = (x_{t+1} - x_t, y_{t+1} - y_t)$, over L consecutive frames are concatenated to form a vector, $\hat{\mathbf{s}} = [\Delta p_t, ..., \Delta p_{t+L-1}]$, which is then normalized to be a trajectory descriptor $\mathbf{s} = \hat{\mathbf{s}} / \sum_{j=t}^{t+L-1} \|\Delta p_j\|$. One drawback of $\mathbf{s}$ is being sensitive to the change of speed over time. In a BoW representation, most code words of $\mathbf{s}$ represent trajectory speed variations, rather than different shapes (Fig. 2 (a)).

Trajectory aligned descriptors In addition, following the procedure in [4], the video volume of a neighborhood of each trajectory is aligned and the resultant volume is described using motion boundary histogram (MBH) [21] and histogram of oriented gradients (HOG) [22].

2.2. OCIT descriptor

To address the limitation of $\mathbf{s}$ descriptor, we propose a novel orientation-based trajectory descriptor. It is important to retain the overall orientation of the trajectory and be invariant to the angle from which the sports video is captured. Thus, all trajectory segments with negative displacement over x-axis are flipped horizontally. Specifically, we compute the orientation of a trajectory segment at frame t by

$$\alpha_t = \tan^{-1}\left(\frac{y_{t+1} - y_t}{\mid x_{t+1} - x_t \mid}\right). \quad (2)$$

The parameter α_t concisely captures the trajectory shape, but confuses some special trajectories, e.g., an upward zigzag and a 45 degree upward straight trajectory. Therefore, $\Delta\alpha_t = \alpha_{t+1} - \alpha_t$ is also used to describe the trajectory shape. We denote the histogram of α_t and $\Delta\alpha_t$ as $\mathbf{h} = [h_1, ..., h_N]$ and $\mathbf{g} = [g_1, ..., g_{N_\Delta}]$. Since tiny motions result in very small trajectory segments, which are less important in overall shape, the contribution of each trajectory segment is weighted by $\|\Delta p_t\|$, for both $\mathbf{h}$ and $\mathbf{g}$. Thus, $\mathbf{h}$ and $\mathbf{g}$ are defined as,

$$h_n = \sum_{t=1}^{L} \delta_n(\alpha_t) \|\Delta p_t\| \; ; n \in \{1, ..., N\}, \quad (3)$$

$$g_m = \sum_{t=1}^{L} \varphi_m(\Delta\alpha_t) \|\Delta p_t\| \; ; m \in \{1, ..., N_\Delta\}, \quad (4)$$

where δ_n and φ_m are the indicator functions,

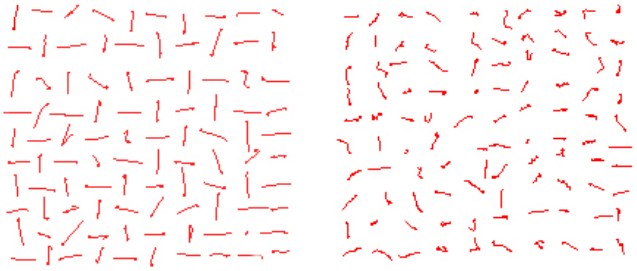
$$\delta_n(\alpha_t) = \begin{cases} 1, & \text{if } \frac{(n-1)\Pi}{N} - \frac{\Pi}{2} < \alpha_t < \frac{n\Pi}{N} - \frac{\Pi}{2} \\ 0, & \text{otherwise} \end{cases},$$

$$\varphi_m(\Delta\alpha_t) = \begin{cases} 1, & \text{if } \frac{2(m-1)\Pi}{N_\Delta} - \Pi < \Delta\alpha_t < \frac{2m\Pi}{N_\Delta} - \Pi \\ 0, & \text{otherwise} \end{cases}.$$

By concatenating $\mathbf{h}$ and $\mathbf{g}$, followed by L_2 normalization, the new trajectory descriptor named Orientation-based Camera angle Invariant Trajectory (OCIT) descriptor, is defined,

$$\text{OCIT} = \frac{(\mathbf{h}, \mathbf{g})}{\sqrt{\| \mathbf{h} \|^2 + \| \mathbf{g} \|^2}}. \quad (5)$$

Given the descriptors computed from a set of training videos, we perform codebook learning via K-means clustering and observe the variability of the code words. As shown in Fig. 2 (a), many of the trajectories for $\mathbf{s}$ look similar, but in fact they represent different paces of movements. Although the training videos contain many trajectories with considerable curvatures, they are not well captured by the code words



(a) (b)

Fig. 2. Representative trajectories of 100 code words for a) $\mathbf{s}$ and b) OCIT. Each trajectory has the minimum distance to one code word.

of $\mathbf{s}$. In contrast, the code words of OCIT (Fig. 2 (b)) capture a vast range of trajectories with different shapes, curvature and overall orientations. Therefore, OCIT is likely to be more effective to represent trajectories than $\mathbf{s}$.

Furthermore, OCIT has a fixed dimension of 15 ($N = 10$, $N_\Delta = 5$), while $\mathbf{s}$ has a dimension of $2L$. Thus, the new trajectory descriptor is both effective and efficient. In the case of limited training data, a lower feature dimension, and hence a lower number of BoW code words, are highly desirable.

2.3. Temporal analysis of videos

Since temporal segmentation is normally not available for sports videos in the wild, it is possible to encounter cases that only a small part of the video contains representative motions. Thus, it is critical to analyze the video in short time segments and properly fuse them to make the final decision based on the most informative segments. Many works split the videos temporally to capture semantics of actions. Some works find the most informative spatio-temporal part of a video using Multiple Instance Learning (MIL) [23, 24], but for our dataset, MIL performed poorly, which is consistent with the finding of [23] that the performance of MIL decreases as the complexity of the datasets increase. In addition, even for temporally segmented videos, there may be ambiguities if the temporal order of trajectories is not taken into consideration.

For TA of motions, each video volume is divided to non-overlapping temporal cells of 1-second length. Histograms of different descriptors based on the trajectories are calculated for each cell. As shown in Fig. 3, the histograms of N_c consecutive cells are then concatenated and L_2 normalized to form the feature representation of one block $\mathbf{b}^k$, where k is the index of the block. Thus, the feature dimension of each block, d, is N_c times the feature dimension of each cell. Blocks slide over cells, with $\frac{100(N_c-1)}{N_c}$ percentage of overlapping between consecutive blocks. Now, in the collection of block features corresponding to a single video, at least one block represents the most informative N_c-second segment of the video. Since TA increases the dimension of video representation by a factor of N_c, we reduce the dimension of $\mathbf{b}^k$ via PCA such that 95% of the variance is retained. If the number of cells in a video is less than N_c, the cells are concatenated and zero padded to form the block.

For a C-category categorization problem, a classifier $f : R^d \rightarrow R^C$ is trained over d-dim block feature $\mathbf{b}^k$ and outputs a L_1 normalized C-dim score vector representing the probability of the block belonging to each of the C categories. Given a test video i, M cells and $M - N_c + 1$ blocks are generated. Experiments show that for blocks where the trajectories are not representative of a specific sports genre, the scores are more randomly distributed over a larger number of categories, and hence the maximum score is relatively low. By denoting the feature representation of block k of video i as $\mathbf{b}_i^k$, and its scores as $\mathbf{x}_i^k (k = 1, ..., M - N_c + 1)$, a weighted fusion is used to compute the final score vector of video i, denoted as $\mathbf{x}_i$ (both $\mathbf{x}_i^k$ and $\mathbf{x}_i$ are C-dim vectors). The weight of each block is the likelihood of the maximum score of the block given a correct categorization, denoted as $p_+(\cdot)$ and estimated by a Gaussian distribution during training. Thus,

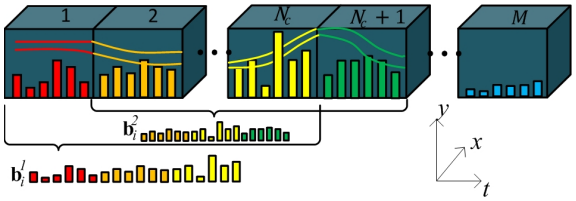


Fig. 3. TA of video blocks, each composed of N_c cells.

$$f(\mathbf{b}_i^k) = \mathbf{x}_i^k = (x_{i,1}^k, ..., x_{i,C}^k) \text{ s.t. } \sum_{c=1}^{C} x_{i,c}^k = 1, \quad (6)$$

$$\mathbf{x}_i = \sum_{k=1}^{M-N_c+1} p_+(\max_c x_{i,c}^k)\mathbf{x}_i^k. \quad (7)$$

The final sports category of video i, y_i, is the category with the maximum value in $\mathbf{x}_i$,

$$y_i = \arg\max_c (x_{i,1}, ..., x_{i,C}). \quad (8)$$

3. EXPERIMENTAL RESULTS

Dataset We collected a dataset of 1,047 videos from 15 sports categories captured by amateur users via a mobile phone app. In each category, 50 videos are used for training and $15 - 25$ videos for testing. The average, max., and min. video length is 35s, 242s, and 1s respectively. Videos are not temporally segmented, so the most informative segment may appear at any part of the video. Our dataset is favorably comparable in size with UCF Sports [25] and Olympic Sports [26] datasets (both are professional sports videos captured by professional TV crew), where each of the three datasets has 15, 9 and 16 categories, and ~ 70, ~ 20 and 50 videos per category respectively. We will make our dataset publicly available[1].

Implementation Details We use K-means clustering to learn a BoW codebook [6] and the implementation in [5] to calculate dense trajectories. To prevent the trajectories of the background or audiences from dominating the trajectories of the players, in all BoW representations the bins with values

[1] http://www.cse.msu.edu/~liuxm/sportsVideo

Table 1. BoW and TA accuracy at various descriptor combinations.

Descriptor combination	BoW	TA
s	38.0%	**43.1%**
OCIT	45.0%	**49.0%**
HOG	55.4%	**61.8%**
HOG+s	57.7%	**64.6%**
HOG+OCIT	60.6%	**67.0%**
MBH	53.3%	**56.9%**
MBH+HOG	65.1%	**66.7%**
MBH+ HOG+s	66.4%	**67.4%**
MBH+ HOG+ OCIT	68.5%	**69.1%**

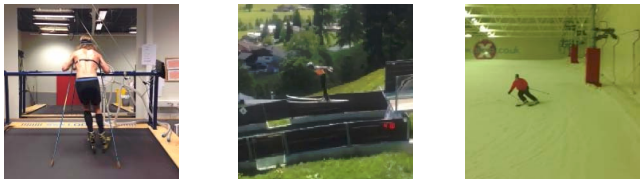


Fig. 4. Sample frames from videos of Skiing category.

larger than $\mu + u\sigma$ are clipped to $\mu + u\sigma$, where μ and σ are the mean and standard deviation of the values in all bins and u is a clipping parameter. This is similar to the clipping normalized histograms in SIFT by a threshold of 0.2 to ensure robustness to illuminations [27]. To compute the trajectories and descriptors, we set $W = 15$, $L = 30$, $N = 10$, $N_\Delta = 5$, $N_c = 5$, and $u = 3$. Default parameters as in [4] are used for trajectory aligned HOG and MBH descriptors. RBF kernel SVM is used for classification and the parameters are tuned via 5-fold cross validation. In all experiments, the number of code words in BoW is 100, 50, 100 and 100 for **s**, OCIT, HOG and MBH respectively. The categorization *accuracy*, the fraction of correctly categorized videos, is used as the metric.

Results of Accuracy As shown in Tab. 1, OCIT outperforms **s** by *itself* (45% vs. 38%) and by *combining* with other descriptors, in both BoW and TA. Considering the compactness of OCIT, this is an impressive result. For *all* combinations of descriptors, TA outperforms BoW. However, as features get richer through combination of descriptors, TA deals with higher dimensionality and results in less improvement due to the curse of dimensionality. Performance of TA for HOG+OCIT is better than BoW for richer combination of MBH+HOG. Note that OCIT is substantially more efficient to compute than MBH. The top performance of **69.1**% is achieved by fusing MBH, HOG and OCIT in the TA approach. To compare with prior work, we implement a state-of-the-art context-based sports categorization method [9], which uses BoW on SIFT descriptors, and receive an accuracy of **42.9**%. This demonstrates that for categorization of sports videos in the wild, motion-based method is more powerful. Figure 4 shows the environment diversity in our data and provide probable reason for the poor result of [9].

Figure 5 illustrates a temporal analysis result of a 31-second video of Hockey category with the label 11. TA extracts 27 blocks for this video. The label assigned to each

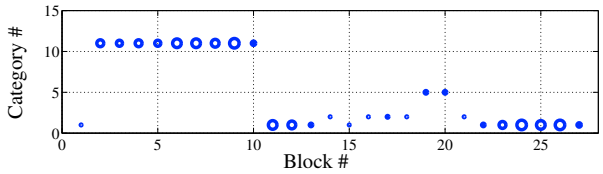


Fig. 5. Labels assigned to temporal blocks of a Hockey video.

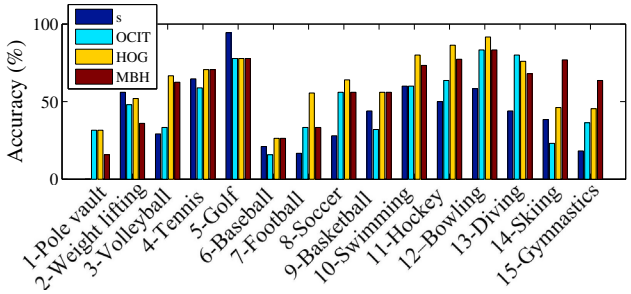


Fig. 6. Categorization accuracy for each category using each single descriptor and the temporal analysis scheme.

block is shown in this figure, with the size of circle representing the weight assigned to each block as in Eqn. 7. The first 10 seconds of the video are more representative of Hockey and are correctly labeled as 11 with larger weights. Therefore, in spite of all the ambiguities in the later part of the video, this video is correctly labeled as Hockey by the temporal analysis.

Figure 6 shows the accuracy of categorization for each category using the TA scheme. Performances are especially low for Pole vault and Baseball. All descriptors confuse Pole vault mainly with Weightlifting. For Baseball, **s**, OCIT, HOG, and MBH mainly confuse this category with Golf, Swimming, Golf, and Volleyball respectively. For categories like Golf and Bowling that have very distinct and limited set of movements, the categorization performance is very good.

Results of Efficiency Since most of the computation time is spent on optical flow calculation, using BoW or TA has negligible computational cost in the test phase. For example, while trajectory and descriptors computation takes $\sim 1,225$s for a 35s-video, BoW and TA for the MBH descriptor take average of 0.002s and 0.018s respectively. Thus, the categorization time using BoW and TA methods is almost the same, and they are on par with other trajectory-based methods [5, 17].

4. CONCLUSIONS

This paper proposes a sports video genre categorization method. We introduce a compact and efficient orientation-based trajectory shape descriptor that is invariant to camera angle variations. A temporal analysis method is introduced to integrate the decisions of local descriptors over time. Superior performance is observed on a dataset of amateur sports videos in the wild when compared to baseline methods.

We plan to extend this approach to a large sports video dataset and human activity datasets. Also, integrating trajectory descriptors with methods in body pose estimation [28,29] can be another future direction.

5. REFERENCES

[1] Jinjun Wang, Changsheng Xu, and Engsiong Chng, "Automatic sports video genre classification using pseudo-2D-HMM," in *ICPR*. IEEE, 2006, vol. 4, pp. 778–781.

[2] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and Jesse S Jin, "A unified framework for semantic shot classification in sports video," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1066–1083, 2005.

[3] Ning Zhang, Ling-Yu Duan, Qingming Huang, Lingfang Li, Wen Gao, and Ling Guan, "Automatic video genre categorization and event detection techniques on large-scale sports data," in *Proc. IBM Conf. Advanced Studies on Collaborative Research*. IBM Corp., 2010, pp. 283–297.

[4] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *CVPR*. IEEE, 2011, pp. 3169–3176.

[5] Heng Wang, Cordelia Schmid, et al., "Action recognition with improved trajectories," in *ICCV*. IEEE, 2013, pp. 3551–3558.

[6] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004, vol. 1, p. 22.

[7] Xun Yuan, Wei Lai, Tao Mei, Xian-Sheng Hua, Xiu-Qing Wu, and Shipeng Li, "Automatic video genre categorization using hierarchical SVM," in *ICIP*. IEEE, 2006, pp. 2905–2908.

[8] C Krishna Mohan and B Yegnanarayana, "Classification of sport videos using edge-based features and autoassociative neural network models," *Signal, Image and Video Processing*, vol. 4, no. 1, pp. 61–73, 2010.

[9] Ning Zhang, Ling-Yu Duan, Lingfang Li, Qingming Huang, Jun Du, Wen Gao, and Ling Guan, "A generic approach for systematic analysis of sports videos," *ACM Trans. Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 46, 2012.

[10] Nazli Ikizler, Ramazan Gokberk Cinbis, Selen Pehlivan, and Pinar Duygulu, "Recognizing actions from still images," in *ICPR*. IEEE, 2008, pp. 1–4.

[11] Weilong Yang, Yang Wang, and Greg Mori, "Recognizing human actions from still images with latent poses," in *CVPR*. IEEE, 2010, pp. 2030–2037.

[12] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid, "Actions in context," in *CVPR*. IEEE, 2009, pp. 2929–2936.

[13] Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge, and Tsuhan Chen, "Spatio-temporal phrases for activity recognition," in *ECCV*. 2012, pp. 707–721, Springer.

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

[15] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld, "Learning realistic human actions from movies," in *CVPR*. IEEE, 2008, pp. 1–8.

[16] Shandong Wu, Omar Oreifej, and Mubarak Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories," in *ICCV*. IEEE, 2011, pp. 1419–1426.

[17] Yu-Gang Jiang, Qi Dai, Xiangyang Xue, Wei Liu, and Chong-Wah Ngo, "Trajectory-based modeling of human actions with motion reference points," in *ECCV*, pp. 425–438. Springer, 2012.

[18] Cen Rao and Mubarak Shah, "View-invariance in action recognition," in *CVPR*. IEEE, 2001, vol. 2, pp. II–316.

[19] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, Cordelia Schmid, et al., "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009, pp. 124.1–124.11.

[20] Martin A Fischler and Robert C Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *ACM Trans. Communications*, vol. 24, no. 6, pp. 381–395, 1981.

[21] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, pp. 428–441. Springer, 2006.

[22] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. IEEE, 2005, vol. 1, pp. 886–893.

[23] Michael Sapienza, Fabio Cuzzolin, and Philip Torr, "Learning discriminative space-time actions from weakly labelled videos," in *BMVC*, 2012, pp. 1–18.

[24] Wei Jiang, Courtenay Cotton, Shih-Fu Chang, Dan Ellis, and Alexander Loui, "Short-term audio-visual atoms for generic video concept classification," in *Proc. ACM conf. on Multimedia*. ACM, 2009, pp. 5–14.

[25] M Sullivan and M Shah, "Action mach: Maximum average correlation height filter for action recognition," in *CVPR*. 2008, pp. 1–8, IEEE.

[26] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *ECCV*. 2010, pp. 392–405, Springer.

[27] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[28] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *CVPR*. IEEE, 2008, pp. 1–8.

[29] Xiaoming Liu, Ting Yu, Thomas Sebastian, and Peter Tu, "Boosted deformable model for human body alignment," in *CVPR*. IEEE, 2008, pp. 1–8.