TRAJECTORY FEATURE FUSION FOR HUMAN ACTION RECOGNITION

Sameh Megrhi, Azeddine Beghdadi and Wided Souidene

L2TI, Institut Galilée, Université Paris 13
99, Avenue Jean-Baptiste Clément, 93430 Villetaneuse, France

ABSTRACT

This paper addresses the problem of human action detection /recognition by investigating interest points (IP) trajectory cues and by reducing undesirable small camera motion. We first detect speed up robust feature (SURF) to segment video into frame volume (FV) that contains small actions. This segmentation relies on IP trajectory tracking. Then, for each FV, we extract optical flow of every detected SURF. Finally, a parametrization of the optical flow leads to displacement segments. These features are concatenated into a trajectory feature in order to describe the trajectory of IP upon a FV. We reduce the impact of camera motion by considering moving IPs beyond a minimum motion angle and by using motion boundary histogram (MBH). Feature-fusion based action recognition is performed to generate robust and discriminative codebook using K-mean clustering. We employ a bag-of-visual-words Support Vector Machine (SVM) approach for the learning /testing step. Through an extensive experimental evaluation carried out on the challenging UCF sports datasets, we show the efficiency of the proposed method by achieving 83.5% of accuracy.

Index Terms— Action recognition, SURF, optical flow, spatio-tempral interest points, frame volume, trajectories, motion boundary histogram.

1. INTRODUCTION

In recent years, the development of video recording technologies, video analysis and processing tools has led to their use in a wide audience in various applications [1, 2]. In order to be efficient and appealing, these new technologies require the implementation of new methods for action /objects recognition. Recognizing human actions from videos is targeted by researchers due to its various applications such as video analysis [3], human-computer interaction [4], surveillance videos [1]. However, action recognition is a challenging task that requires handling occlusions, scale changes, illumination, background clutter, and viewpoint changes (see Figure 1).

In fact, a robust action recognition is based on relevant video description. The state-of-the-art attests a large variety of video descriptors [5]. Among them, spatio-temporal local features have been widely used in order to recognize human actions, objects and events in videos [5]. One of the proposed spatio-temporal descriptors is the ST-SURF proposed by [6]. In addition to the spatial information captured by the SURF descriptor, ST-SURF globes motion and localization information. The latter contributes to inject spatial information recommended especially when using Bag-of-Words in the classification level. A fusion step of SURF and trajectory cues leads to a video descriptor. The extracted feature contains spatial and temporal, trajectory and motion information. We propose in this paper a video description based on an optimized spatio-temporal features called OST-SURF.

(a) Illumination. (b) Scale change. (c) Background clutter.

Fig. 1: Some challenges of action recognition

In order to reduce the computational load, authors in [6] proposed a temporal video segmentation in frame packets and demonstrated that this segmentation is simple yet effective. The main idea behind this segmentation is to detect SURF descriptors, proposed by [7], and to track their displacements in every frame.
In this paper, we build on their insight and optimize their method by setting a minimum displacement angle to discriminate relevant actions from camera motion which are usually small displacements. In addition, we propose to reduce small camera motion by using the Motion Boundary Histogram (MBH) . MBH descriptor produces interesting recognition

results especially in videos containing camera motion [5]. The latter is extracted from a parametrization of the optical flow fields. We finish by the extraction of an optimized spatio-temporal descriptor called OST-SURF which adds spatial information to the final descriptor.

The final descriptor is used in a bag of visual words (BoVW) representation. The latter can be introduced in a training/testing process to make the difference between action's classes. In this paper we employ a K-means clustering algorithm to quantize the extracted descriptors. Yet, each video clip is represented by a histogram of K-bins. We experimentally show that this contribution outperforms other state of the art approaches on a complex and realistic dataset [8]. The rest of the paper is structured as follows: Section 2 is a review of the related works. Section 3 is a detailed description of the proposed video segmentation approach. In section 4 the extracted features are detailed. In section 5 the experimental settings and evaluation results are reported and discussed. Finally, the conclusion is drawn in Section 6.

2. RELATED WORK

To track human action in videos, many recent researches focus on extracting relevant descriptors from a fixed frame number [9]. For instance, Noguchi et al. [3] chose to divide video sequences into snippets of five frames. Skindler et al. [10] suggested that action recognition systems require one to 10 frames to ensure good performances. In [11], it has been shown that a fixed frame number can be exploited in video containing periodic unique actions with static background. That is why, authors in [6], introduce a video segmentation into frame packets based on the trajectory tracking. In this case, authors perform video segmentation based on the SURF's motion trajectory tracking. The size of the segmented packets is not fixed since it depends on the detected action. In this paper, we exploit their method and improve the segmentation process by adding a motion angle threshold to discard small motion. We obtain video segments containing relevant actions.

In order to extract the spatio-temporal features, one of the most widely used methods are local cuboids [3]. Indeed, Dollar et al. [12] and Laptev et al. [13] extract histogram of gradient (HoG) and histogram of flow (HoF) respectively of a cuboid. These descriptors attest a good accuracy. However the proposed techniques faced many issues [3]. First, they are time consuming. Second, it is hard to set the cuboid size. Finally, the authors join temporal and spatial patterns into a common 3D space which recently suffers from many limitations [5]. For the aforementioned reasons, several recent researches focus on detecting IPs and tracking them through time to extend them to the temporal domain. Recently, researchers target the tracking of the motion of IPs. This allows exploring several motion cues such as velocity [14], location [15], trajectory curves [16] or different motion cues combinations [6]. The trajectories can be extracted by matching interest points [17], or by using a tracker such as KLT (Kanade-Lucas-Tomasi) tracker [18] which is used to extract trajectories in videos, or practical filter tracking schemes [19]. Recently, Wang et al. proposed Dense Trajectory tracking to encode temporal information [4]. They demonstrate that trajectory tracking is an intuitive and successful approach in several public benchmarks.

The task of extracting robust features to moving camera and a dynamic background is very challenging. Although, many schemes have been proposed to reduce small camera motion [20]. Our goal is to develop a video presentation which discards small camera motion without sacrificing significant human action cues. To this end, we included in our proposed scheme the MBH feature proposed by [21]. MBH descriptor has been extracted from the gradient of optical flow. It removes constant motion and preserve significant motion. MBH was employed in various action recognition schemes [5]. MBH is not dedicated to remove camera motion, but combined with the OST-SURF, it will contribute significantly to camera motion compensation. The OST-SURF is a spatio-temporal SURF obtained by the tracking of the detected SURF points. The particularity of this descriptor is that, it is not only compact and reliable but also it focuses only on moving objects.

3. VIDEO TEMPORAL SEGMENTATION

In the following, we recall some ideas on which our proposal is based. Authors of [6], start by detecting SURF descriptors, and then cluster these IPs into groups. For every moving IPs between the frames (n-1, n and n+1), they compute an angle called α. The latter is between the lines supporting the motion of an IP from the couple of frames (n-1, n) and (n, n+1). For every cluster of IPs, they extract an average angle α_{avg} and compare it to α_{max} (a parameter fixed empirically). They finally segment a succession of frames, that they call FV, in which each group of IP has an α_{avg} lower than α_{max}. See Figure 2 for details. In order to optimize this technique, we introduce a third parameter called α_{min}. The role of this angle is to allow discarding the small motion as illustrated in Figure 3.

The angle α_{avg} must be greater than α_{min} and less than α_{max}. Thus, we discard small undesirable camera motion. The major steps of the resulting segmentation algorithm are shown in Table 1.

4. DESCRIPTOR EXTRACTION

In this work, video description is performed through a late-fusion process of two descriptors that we describe in this section.

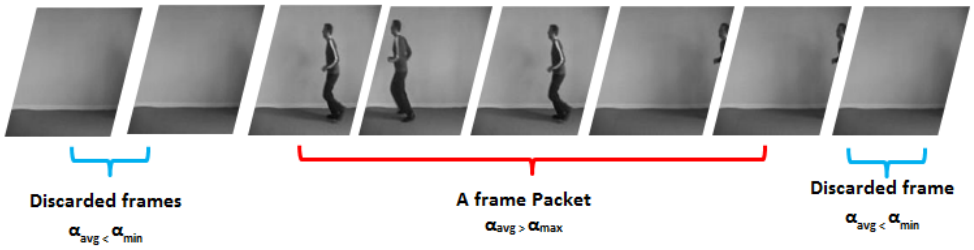

Fig. 3: Proposed FPs segmentation.

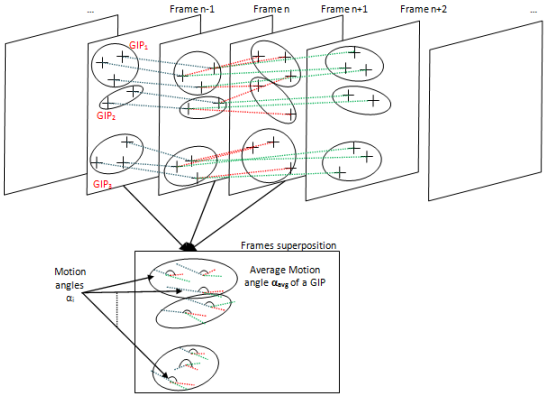

Fig. 2: IPs trajectory tracking for FPs segmentation.

Table 1: Proposed algorithm.

Input : I - input video;
$\alpha_{min}, \alpha_{max}$ - motion angles;
Algorithm :
step1 IP extraction from frames $\{f_1, f_2\}$;
step2 Groups of IPs defined;
step3 Compute the line supporting the motion;
Apply the above three steps to $\{f_2, f_3\}$;
Compute the angle between each motion line;
Extract α_{avg} for each GIP;
if $\alpha_{avg} \leq \alpha_{min}$;
then go to the next frame;
else Compare α_{avg} to α_{max};
end if
repeat previous steps;
until $\alpha_{avg} \geq \alpha_{max}$;
Output = f_n, t_{min}, t_{max} ;

4.1. SURF Tracking and ST-SURF Extraction

ST-SURF was introduced by [6]. The latter captures spatial and temporal information. The main idea is to extract the trajectory of a SURF point by tracking its motion trajectory. The authors used Hessian Matrix to detect salient points. Features' tracking is based on the optical flow. They employed the Sun and al. [22] optical flow computation algorithm. In fact, Sun and al. proposed a median filtering to denoise the optical flow, exploiting connections between median filtering and ℓ_1 based denoising. They proved that algorithms relying on a median filtering step approximately optimizes a different objective that regularizes the flow over a large spatial neighborhood [22]. The resulting algorithm ranks first in both angular and end-point errors in the Middlebury evaluation [22]. In this paper, every Fv consists on a flexible frame number f_n. We assume that every f_n corresponds to a volume of frames in the 3D space. This cubic volume is characterized by a frame number f_n from t_{min} to t_{max}, its frames' surfaces dimensions FS and its center FVc. In the first frame of a video sequence, a given $IP = (x, y, 1)$ is defined by its position (x, y) and its frame number 1. In the frame $(1 + n)$, the IP moves by a displacements u in the x direction, and v in the y direction. IP becomes, $IP(1 + n) = (x + u, y + v, 1 + n)$. Thus, in the FPV the 3D direction (u, v, n) represents the direction of the IP motion. The vectors (u, v) are computed by the approach described in [22]. Figure 4, illustrates the cube and its projection into the planes (t, x) and (t, y).

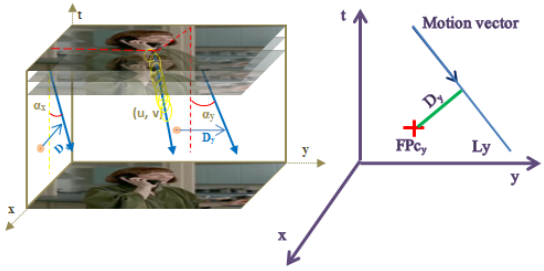

Fig. 4: The projection of a motion vector in the adjacent planes.

We project the motion vectors onto the planes (t, x) and (t, y) of the FPV to define an angle for each projection. The first angle α_x between optical flow and the plane (t, x), the angle α_y between the plane (t, y) and the motion vector.

$$\alpha_x = 90 - \frac{180}{\Pi} arctan(u/f_n). \quad (1)$$

$$\alpha_y = 90 - \frac{180}{\Pi} arctan(v/f_n). \quad (2)$$

For each IP, we project its motion vector onto the planes (t, x) and (t, y) and obtain two lines L_x and L_y. The orthogonal projection of FVc_x and FVc_y onto the lines L_x and L_y allows the computation of both distances D_x and D_y between the cube center and the lines supporting the motion vectors L_x and L_y.
For an IP located at (x, y, t), D_x and D_y are given by:

$$D_x = D_{xu} - D_{tv}, D_y = D_{yv} - D_{tu}, \quad (3)$$

where

$$D_{xu} = (x - x_{max}/2)cos(180/\Pi \times arctan(u/f_n)), \quad (4)$$

$$D_{tv} = (t - t_{max}/2)sin(180/\Pi \times arctan(v/f_n)), \quad (5)$$

$$D_{yv} = (y - y_{max}/2)cos(180/\Pi \times arctan(v/f_n)), \quad (6)$$

$$D_{tu} = (t - t_{max}/2)sin(180/\Pi \times arctan(u/f_n)). \quad (7)$$

4.1.1. ST-SURF Extraction:

An early fusion is performed to generate the OST-SURF. In fact, we concatenate the 64-D SURF feature vector and the temporal 4-D feature. The resulting descriptor is a 68-D spatio-temporal descriptor vector. This allows extending the spatial SURF descriptor to videos. These features track the interest point through time in each FV. In our work, we consider only moving interest points (where $\alpha_x \neq 0$ and $\alpha_y \neq 0$). Our contribution in the feature extraction level consists on the introduction of a new parameter called frame number f_n.

4.2. Motion boundary histogram (MBH)

The motion boundary histogram (MBH) was introduced in [21] to detect action. MBH computes the gradient of the optical flow fields in both (x, t) and (y, t) directions. Hence, it captures salient optical flow changes while suppressing small motion. The latter is usually derived from camera motion. The final MBH_x and MBH_y are 96-D $(2 \times 2 \times 3 \times 8)$ features vector. In this work, we used MBH, not only for its ability to reduce camera motion, but also as an efficient motion descriptor [21, 5].

4.2.1. Descriptors Learning/Evaluation Pipeline:

After extracting ST-SURF and MBH descriptors, we construct a BOvW separately for every descriptor. The basic idea is to assign a set of objects into groups so that the objects of similar type will be in one cluster in order to construct a visual codebook. The latter can be used to represent an action, a scene or an object. The generated descriptors are quantized into visual words using k-means clustering. Each video sequence can be then represented as the frequency histogram over the visual words. The resulting histograms of visual word are used as inputs to the classification process.

5. EXPERIMENTS

In the following, we describe the dataset used in the evaluation of the proposed work. We evaluate the proposed descriptors in a bag-of-features based action classification task and compare our approach to the state-of-the-art methods given in [23].

5.1. Experimental Setups and Data

5.1.1. Dataset:

The proposed framework is tested on UCF sports dataset [8]. This dataset is a realistic and challenging data obtained from broadcast sport videos by Ahmed et al. [8]. The collection represents a natural pool of actions featured in a wide range of scenes and viewpoints. The publicly available part of this dataset contains nine actions namely diving, golf, swinging, kicking, lifting, horseback riding, running, skating, swinging and walking. This dataset contains 200 video sequences with a resolution of 720 $\times$480.

5.1.2. Parameter Settings:

In our experiments, we explored optimal parameter settings proposed by the state-of-the-art [23]. The empirically optimal size book is k = 4000 with α_{max} = 42°, α_{min} = 4° and group of IP = 38. These settings gave us satisfactory results in term of accuracy.

5.2. Experimental results

On Table 2, the first row reports the Best Accuracy (**BA**) for the different detector/descriptor combinations reported by other works on the UCF sports dataset [24]. The Average Accuracy for Hessian (**HA**) detector/descriptor combinations on the UCF sports dataset are drawn in the second row. In a dense sampling, Wang et al. achieved an accuracy of 77.4% using the HoG descriptor and 82.6% using the HoF descriptor. Indeed descriptors of local motion, given by the histograms of optical flow (HoF) characterize the action better than the histograms of oriented gradient (HoG) that describe the local appearance. They also obtained an accuracy of 81.6% using the HoG/HoF combination, the result of action recognition is not improved because the HoG are less accurate to characterize temporal information. The extension of HoG in the time domain, associated with Gabor detector, allowed

Klaser et al. to reach 85% using HoG3D/Gabor. The spatial orientation of this feature describes the information appearance. The temporal orientation extracted describes movement speed. Using Hessian detector, the combination HOG/HOF outperforms the HOG, HOF and the HOG3D. This underlines the importance of the choice of the IPs detector. The proposed optimized version of ST-SURF achieves 83.5% of accuracy outperforming the ST-SURF by 2.8%. This is due to many reasons. In fact, the optimization of the video segmentation boosts significantly the accuracy of the action recognition. MBH is a relevant descriptor which exploits motion information and shows its robustness in realistic video. This proves the importance of camera motion reduction in action recognition.

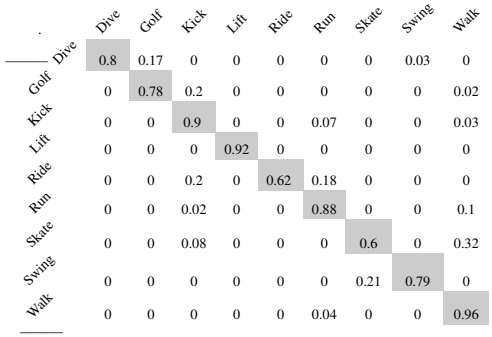

Fig. 5: Confusion matrix of the classification results for the UCF sport dataset using ST-SURF descriptor.

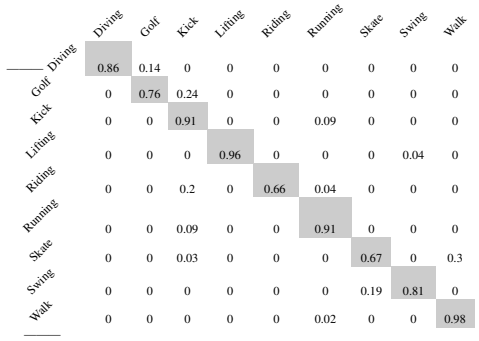

Fig. 6: Confusion matrix of the classification results for the UCF sport dataset for the proposed approach using the combination of optimized ST-SURF and MBH descriptors.

Figure 5 is the confusion matrix that describes the classification result for the ST-SURF feature reported from [6]. We emphasize that the lowest results are reported in the "skate" and "ride" actions because the movements of these actions are horizontal. The accuracy is improved gradually as the actions contain vertical movements as "walk", "kick" and "lift" where we see a high rotation. This proves that ST-SURF is not robust to linear horizontal motion. Figure 6 is the confusion matrix that describes the classification result for the proposed feature. The latter is the Fusion of OST-SURF and the MBH features. The reported results prove that the proposed approach improves the accuracy and classification in several actions such as "diving", "Golf" and "kicking". If we focus on the "golf" action, in Figure 5, ST-SURF detects "golf" by 0.9%. However it confuses "golf" action with other actions such as "running" 0.07% and walking 0.03%. The proposed video description, in the same settings, reduces the confusion and achieves 0.91% for "golf" and 0.09% for "running" eliminating the "walking" confusion. The accuracy of "dive" and "swing" actions are close due to the similarity between these two actions. Less accuracy is observed in the "jogging" and "running" actions because these actions are also almost similar. Regarding all these results our method is equivalent, and even better in same cases, to the state-of-the-art.

6. CONCLUSION

In this paper, we proposed a novel scheme to segment video into a discriminative action sequence. To this end, we also proposed an optimized spatio-temporal descriptor based on spatio-temporal interest points. The proposed feature extraction consists in detecting IPs in a frame volume and mapping them into the temporal domain based on the optical flow orientation and position. To improve action detection, we explored motion boundary histogram descriptor. In one hand, MBH is a relevant motion descriptor. In another hand, it contributes to remove small camera motion effects. It is shown that by using the late fusion a robust high level video description is obtained. The proposed framework demonstrates promising recognition performance achieving the accuracy score of 83.5% in UCF sports. The obtained results proves that we are already equivalent to the state-of-the-art performances. In future work, we plan to investigate different interest points detectors as well as larger and more realistic dataset in a bag-of-visual-words based representations.

7. REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *Transactions onSystems, Man, and Cybernetics*, 2004.

[2] PLM. Bouttefroy, A. Bouzerdoum, S. Phung, and A. Beghdadi, "Abnormal behavior detection using a multi-modal stochastic learning approach," in *International Conference on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 2008.

[3] A. Noguchi and K. Yanai, "A surf-based spatio-temporal feature for feature-fusion-based action recognition," in *Trends and Topics in Computer Vision*. 2012.

[4] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *International Con-*

Table 2: Average accuracy for various detector/descriptor combinations on the UCF sports dataset.

Descriptor \ Accuracy	HOG3D	HOG/HOF	HOG	HOF	E-SURF	ST-SURF	Proposed
BA	85%	81.6%	77.4%	82.6%	77.3%	80.7%	**83.5%**
HA	78.9%	79.3%	66.0%	75.3%	77.3%	80.7%	**83.5%**

ference on Computer Vision and Pattern Recognition, 2011.

[5] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, 2013.

[6] S. Megrhi, A. Beghdadi, and W. Souidene, "Spatio-temporal surf for human action recognition," in *Advances in Multimedia Information Processing, PCM 2013*. Springer, 2013.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European Conference on Computer Vision*. 2006.

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

[9] T. Guha and R.K. Ward, "Learning sparse representations for human action recognition," *Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[10] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?," in *Conference on Computer Vision and Pattern Recognition*, 2008.

[11] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *International Conference on Pattern Recognition*, 2004.

[12] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.*, 2005.

[13] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition," in *Spatial Coherence for Visual Motion Analysis*. 2006.

[14] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *International Conference on Computer Vision*, 2009.

[15] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion models," *Advances in Neural Information Processing Systems*, 2003.

[16] C. Rao, A. Yilmaz, and M. Shah, "View-invariant representation and recognition of actions," *International Journal of Computer Vision*, 2002.

[17] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, 2004.

[18] B.D. Lucas, T. Kanade, et al., "An iterative image registration technique with an application to stereo vision.," in *International Joint Conference on Artificial Intelligence*, 1981.

[19] PLM. Bouttefroy, A. Bouzerdoum, S. Phung, and A. Beghdadi, "Vehicle tracking using projective particle filter," in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance.* IEEE, 2009.

[20] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision*. 2012.

[21] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*. 2006.

[22] D. Sun, S. Roth, and M.J. Black, "Secrets of optical flow estimation and their principles," in *International Conference on Computer Vision and Pattern Recognition*, 2010.

[23] A. Kläser, Heng W., and M. M. Ullah, "Evaluation of local features for action recognition @ONLINE," 2009.

[24] H. Wang, M. M. Ullah, A.r Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009.