Action Recognition: A Region Based Approach

Hakan Bilen[1], Vinay P. Namboodiri [1] and Luc Van Gool [1,2] *
[1]ESAT-PSI/IBBT, K.U. Leuven, Belgium
[2] Computer Vision Laboratory, BIWI/ETH Zurich, Switzerland
firstname.lastname@esat.kuleuven.be

Abstract

We address the problem of recognizing actions in real-life videos. Space-time interest point-based approaches have been widely prevalent towards solving this problem. In contrast, more spatially extended features such as regions have not been so popular. The reason is, any local region based approach requires the motion flow information for a specific region to be collated temporally. This is challenging as the local regions are deformable and not well delineated from the surroundings. In this paper we address this issue by using robust tracking of regions and we show that it is possible to obtain region descriptors for classification of actions. This paper lays the groundwork for further investigation into region based approaches.

Through this paper we make the following contributions a) We advocate identification of salient regions based on motion segmentation b) We adopt a state-of-the art tracker for robust tracking of the identified regions rather than using isolated space-time blocks c) We propose optical flow based region descriptors to encode the extracted trajectories in piece-wise blocks. We demonstrate the performance of our system on real-world data sets.

1. Introduction

Given a video clip, humans have no problem in understanding the actions happening in the video. This is so irrespective of the setting of the scene, the persons involved or the viewpoints. Humans are also able to understand a wide variety of actions ranging from simple actions such as a person standing up or sitting down, to complexer sequences of actions like cooking. Achieving the same computationally is a very challenging problem.

All current approaches try to solve this problem by describing the motion information in the scene. The features used to describe the motion information are crucial. There has been significant progress lately by methods that extract space-time interest point features [12, 13]. However, we believe that it is worthwhile to explore the use of more generally shaped spatio-temporal regions for describing the motion in the scene. There have been similar investigations done by the community for recognizing objects in an image [9, 1]. The use of local regions for describing motion is however not common in action recognition. This is, because a local region to describe motion is not as well defined as a space-time interest point. One could use motion segmentation to obtain an initial segmentation for a frame. However, describing a segment requires it to be temporally connected as well. Tracking segments based on individual frame-wise motion segmentation is difficult, as any specific segment may deform rapidly. In this paper, we address this issue and show that using a state-of-the art robust region tracking algorithm [3] one can now obtain stable region descriptors to describe motion information.

1.1. Overview

An overview of our approach is given in Fig. 1. Given a training video of an action, we apply motion segmentation [5] to obtain connected components that correspond to moving regions. Then, we use a state-of-the-art tracker [3] that can robustly track arbitrary regions using pixel-wise posteriors. Local shape changes of the regions are handled by variational level sets. This robust tracking of regions enables us to obtain stable motion flow patterns. We compute the optical flow in these regions and calculate the histograms of orientation of the optical flow. During the training phase, we cluster the histograms from all training videos using standard k-means to obtain a vocabulary of optical flow. Subsequently, we represent each training video as a set of vector-quantized motion histograms. Finally, we train an SVM classifier on the training data. Using that classifier we can assign action labels to new, previously unseen videos that have been subjected to the same feature extraction and quantization stages.

*Work is funded by IBBT iCocoon

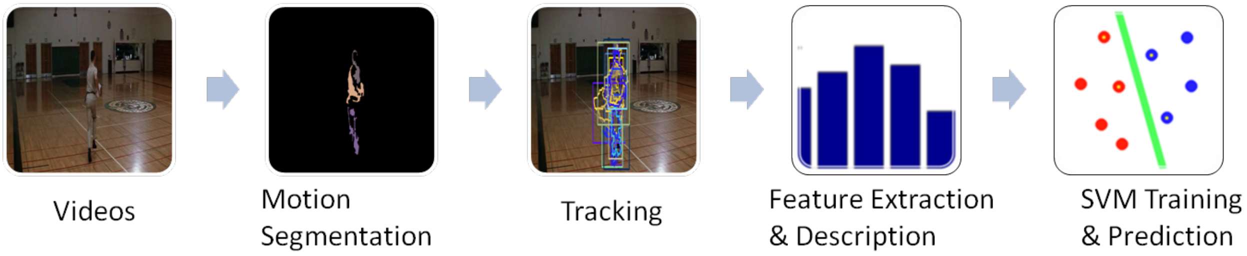


Figure 1. Overview of the proposed approach.

2. Related Work

There has been considerable interest towards addressing and understanding human actions. Initially, the methods for addressing the problem of action recognition were based on global representations of human motion. Among these are the early works by Bobick and Davis [4] that captured view dependent motion and the parameterized motion models by Yacoob and Black [20]. Ali *et al.* [2] proposed kinematic flow features to recognize actions. Another approach has been correlation based human motion classification by Efros *et al.* [7] and Zelnik-Manor and Irani [22]. Lately this approach has been pursued by Shechtman and Irani [16] where they match self-similarities across images and videos.

There has been more widespread interest lately in methods that use space-time interest points (STIP). Initial work towards using these features was by Schuldt *et al.* [15] where the authors used features extracted using the framework of Laptev and Lindeberg [11]. Subsequently there have been other a large number of variants of space time interest points. Dollár *et al.* [6] extracted cuboids with a detector using a 2D Gaussian kernel in spatial domain and Gabor quadrature filters temporally. Kläser *et al.* [10] proposed spatio-temporal descriptor based on 3D gradients. In the work by Wang *et al.*[17], the authors comprehensively evaluate the various spatio-temporal descriptors and also propose uniform dense extraction of features. In the work by Willems *et al.* [18], the authors propose an exemplar based method for action recognition where they use an extended SURF [19] based spatio-temporal detector and also obtain probable hypotheses to localize the action. Laptev *et al.* [12] proposed a data set extracted from Hollywood movies. This data set was considerably more challenging than previously used simple data sets as it contains a lot more variation of viewpoints, background clutter and scenarios. Marszalek *et al.*[13] extended that data set further and use contextual information in their approach. Lately, Yeffet and Wolf [21] have used a spatio-temporal extension of local binary patterns for action recognition and a recent work by Messing *et al.* [14] explicitly tracks interest point features for action classification.

Our approach uses local information, too, but differs from the methods above in that we use a region based feature description. Further, we also handle the features rather differently. Rather than considering isolated spatio-temporal regions we are interested more in the temporal continuity of regions.

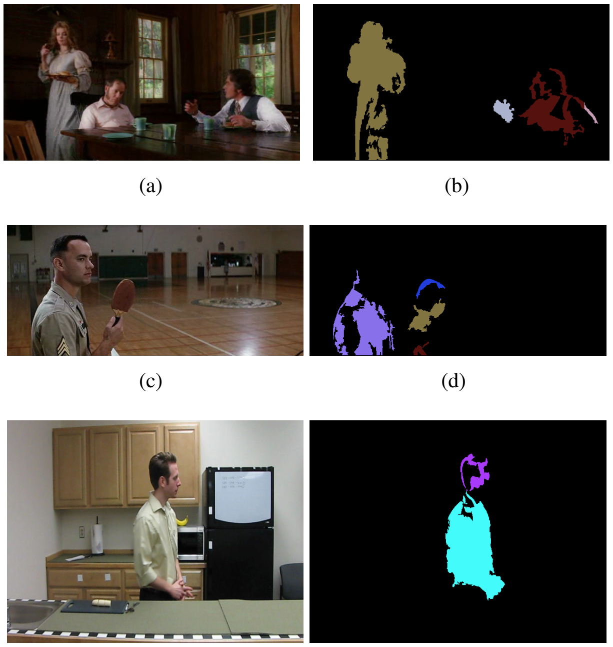
3. Action Representation

We use a bag of words classification system for classifying action [12]. There are three steps towards obtaining the features for this representation. We first detect salient motion regions, we track these regions, and we then describe these regions with an optical flow descriptor.

3.1. Detecting Salient Regions

In order to detect the salient motion regions we employ a simple motion segmentation algorithm [5]. This segmentation algorithm uses a temporal motion history image and is computationally fast. The segmentation algorithm uses frame differencing to identify the most recent silhouette and extends it using a flood fill algorithm to connect areas of motion. The advantage of this simple approach is that it enables us to quickly identify the regions of interest. This algorithm provides us with a coarse region mask, i.e. the detected region includes some background. These regions are subsequently refined in our framework using level-set based segmentation and tracking described in 3.2, where the actual region used shrinks to the correct object boundary.

We eliminate all regions with an area less than 7 pixels and motion magnitude less than 2 pixels. We show examples of this motion segmentation in Fig. 2. There are cases when this algorithm fails to segment the motion. This does not have an adverse effect as we are continuously tracking each detected region. As indicated in the examples, the regions provided by motion segmentation coarsely identify the approximate location where action is happening by various patches. We perform motion segmentation in each frame. In case some region is not identified correctly in one frame, it will most likely be segmented in some of the following frames. As a result, we obtain an over segmentation of the motion region. This is desirable in order to compre-



(a) (b) (c) (d) (e) (f)

Figure 2. Motion Segmentation Examples. The left column (a,c,e) shows the frames and right column (b,d,f) shows motion segmentation examples corresponding to these frames.

hensively describe the motion region. In order to partially alleviate the redundancy in segments we do pruning of these regions during the tracking phase.

3.2. Robust region tracking

Our tracking method is based on a recently published method for robust visual tracking using pixel-wise posteriors [3]. The posterior based tracker incorporates non-parametric distributions into a region-based level-set function that allows it the online adaptation of the region's appearance model. The appearance model is modeled by a foreground model M_f and a background model M_b that are HSV histograms of the region and its immediate surrounding background. The foreground regions are obtained by motion segmentation as described above. The foreground region is inflated by a factor s and its difference to the original foreground region is taken as background region.

The algorithm includes three main steps, a rigid registration between the frames through a warp W, segmentation of the shape kernel $\Phi(\mathbf{x})$ and appearance learning. The registration computes the rigid motion, segmentation handles the shape deformations and appearance learning updates the foreground and the background models. During tracking, the pixels inside and outside the computed contour are used to update those histograms online. The foreground and background models M_f, M_b are marginalized and the pixel-wise posterior probability of the shape Φ and the position parameter vector p given the pixel location $\mathbf{x}$ and values $\mathbf{y}$ in the object frame are obtained. The pixel-wise posteriors $P(M|\mathbf{y}_i)$ are fused to update the shape and position of the tracker in each frame:

$$P(\Phi, p|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} \left\{ \sum_{M} P(\mathbf{x}_i|\Phi, p, M) P(M|\mathbf{y}_i) \right\} \cdot P(\Phi)P(p) \tag{1}$$

where N denotes the total number of pixels in the region.

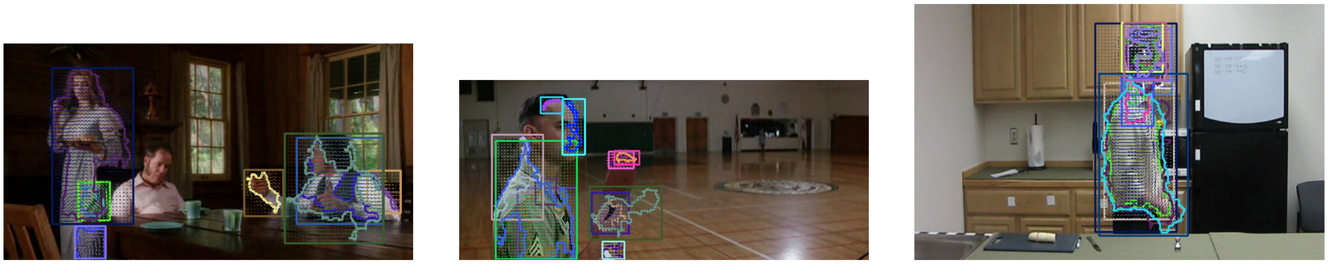
Instead of solving the problem directly in a segmentation framework, the authors [3] decompose the tracking into two steps, a rigid registration and segmentation. The segmentation is achieved by evolving the level set using calculus of variations. This is done by carrying out gradient ascent using the gradient flow. As a result, any extraneous background pixels in the motion segmented region are removed. Incorporating a warp $W(x, \Delta p)$ into eqn. 1 and taking logs the resultant cost function is then given by eqn. 2 which can be further optimized to compute frame-to-frame rigid transformation:

$$\log(P(\Phi, p|\mathbf{x}, \mathbf{y})) = \sum_{i=1}^{N} \{\log P(W(x_i, \Delta p)|\Phi, p, y_i)\} \cdot P(\Phi)P(p) \tag{2}$$

The authors [3] suggest approximating the warp by a first order Taylor series expansion that suffices for rigid registration. Having registered the embedding function first by using optimization result from eqn. 2, local shape changes are handled by evolving the variational level-set $P(\Phi, p|\mathbf{x}, \mathbf{y})$. Further details are explained in the paper [3]. The tracked regions for a few sample frames are shown in Fig. 3. The transformed region Φ is compared with the new motion segments obtained by the segmentation routine. If the overlap between the regions indicates that they are similar then the new motion segment is pruned. Through this method for tracking, we are able to track arbitrarily shaped regions and obtain longer and more stable trajectories.

3.3. Describing Regions

We describe each trajectory using contiguous optical flow descriptors. The process of computing the descriptor is illustrated in Fig. 4. For each frame we compute the optical flow using the two-frame motion estimation algorithm [8]. The magnitude and direction of the resulting optical flow field for each region is shown in Fig. 3(a,b,c). We then compute the histogram from the orientation of optical flow for a specific region bounded by the region boundary Φ. We use sixteen bins for the histogram of optical flow.



(a) (b) (c)

Figure 3. Tracking Examples. The figures show different frames to illustrate the regions tracked. The direction and magnitude of the optical flow are also indicated in these frames.

In the next stage, in order to capture the varying motion flow pattern over time, we accumulate the optical flow over different temporal windows. We specifically use number of frames $\tau = \{1, 2, 4, 8\}$. For the i^{th} frame we add the histogram from the current frame and the following $\tau - 1$ frames and obtain a cumulative histogram of optical flow. This histogram is normalized.

The final descriptor is obtained by concatenating the histogram for the different τ values. The resulting descriptor contains 64 values. There are significant differences in the region descriptor proposed here as compared to the HoF descriptor [12] used to describe space-time interest points. In this descriptor, we specifically restrict the optical flow computation spatially to the region of interest. This descriptor is captured along the trajectory of motion. Therefore, even temporally, the region described is the same. This is different from the usual axis aligned cuboid used to capture the motion information around an interest point. Since a region is more spatially extended than a point, an axis aligned cuboid around a region would capture far too much extraneous information. Thus by using the trajectory we are able to obtain a more coherent region descriptor.

This descriptor is computed for the trajectory at each frame. We explored use of a stepping interval to skip frames but found that computing the descriptor for each frame worked better. We intend in future to further evaluate various modifications of the proposed region descriptor. Note that each descriptor individually describes a region at a frame. It would be interesting to consider grouping of different regions belonging to same object. This is however not explored currently.

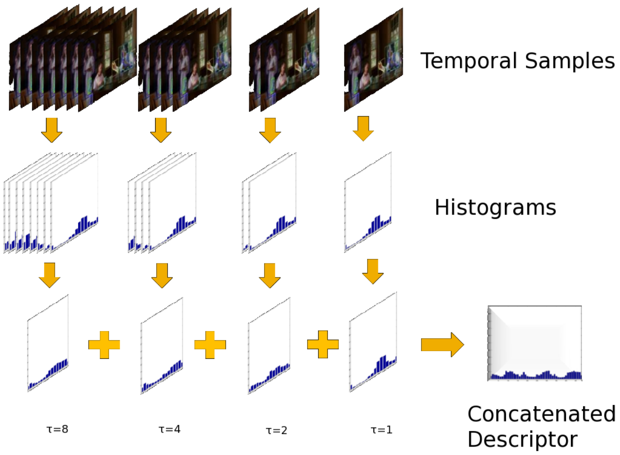


Figure 4. Illustration of the process of computing the region descriptor.

4. Learning feature sets

In order to learn the features we follow the framework used commonly in action recognition [12, 13] based on bag of features. We specifically follow the setting described by Wang *et al*. [17]. Given a training set, we extract optical flow descriptors from trajectories. We then randomly sample 100,000 features from the training set and cluster them with k-means to obtain a fixed spatio-temporal feature vocabulary of 4000 words. The vocabulary is then used to quantize the features and obtain a histogram of these features. This is the feature vector used for classification.

We classify the feature vectors using a non-linear support vector machine with a kernel based on χ^2-distance. The distance between two encoded feature sets is the χ^2-distance between the two feature vectors given by

$$D_c(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}} \tag{3}$$

where h_{in} and h_{jn} are the individual histogram bins of the histograms H_i and H_j respectively. The kernel on the histograms is then given by the following Gaussian kernel

$$K(C_i, C_j) = \exp\left(-\frac{D_c(H_i, H_j)}{\gamma}\right) \tag{4}$$

Here, C_i, C_j are the feature vectors corresponding to video clips i and j respectively. We obtain the parameter γ through cross-validation on the training data set.

(a) AnswerPhone (b) ChopBanana (c) DrinkWater (d) UseSilverWare (e) WriteOnWhiteBoard

(f) AnswerPhone (g) GetOutCar (h) HandShake (i) HugPerson (j) Kiss

Figure 5. Sample frames from *daily living* (top row) and *Hollywood2* (bottom row) data sets.

	answerPhone	chopBanana	dialPhone	drinkWater	eatBanana	eatSnack	lookUpInPB	peelBanana	useSilverware	writeOnWB
answerPhone	1.00	0	0	0	0	0	0	0	0	0
chopBanana	0.067	0.866	0	0	0	0	0	0.067	0	0
dialPhone	0.133	0.067	0.733	0	0.067	0	0	0	0	0
drinkWater	0	0	0	0.867	0	0	0	0.133	0	0
eatBanana	0.133	0.133	0.267	0	0.470	0	0	0	0	0
eatSnack	0	0	0	0	0	0.733	0.133	0	0.133	0
lookUpInPB	0	0	0	0	0	0.067	0.733	0.067	0.133	0
peelBanana	0	0	0.267	0.067	0.067	0	0	0.470	0.133	0
useSilverware	0	0.067	0	0.133	0	0	0.133	0	0.670	0
writeOnWB	0.133	0	0	0	0	0	0	0	0	0.867

Table 1. Confusion matrix for the activities of the *daily living* data set.

5. Experimental Evaluation

We have experimentally evaluated our method on two different challenging data sets which we obtained from the authors' websites. Foremost we considered the activities of *daily living* data set [14]. This contains ten different types of complex actions (like looking up a phone number in a telephone directory, and eating food with silverware). These activities are performed three times by five people with different sizes, genders, shapes and ethnicities. Videos are taken at high resolution (1280 × 720 pixels). A few samples of the actions are shown in Fig. 5.

Next we consider the performance of our method on the *Hollywood2* data set [13]. This data set is comprised of 12 actions (like answering the phone, driving car, eating, etc.) that have to be recognized from real-world challenging movie sequences. Some examples of the actions and the clips are shown in Fig. 5. In our experiments, we used the same data set setting as used by Wang *et al.* [17] in their evaluation.

As explained in section 4, we obtain feature vectors using a fixed vocabulary size of 4000 words. We evaluate the results using the standard criterion as used in the original data set evaluation.

The *daily living* data set [14] contains 150 videos of 10 different actions performed by 5 different persons. The action recognition is done by training on 4 persons and testing on the remaining person. This leave-one-out strategy is used on all subjects and the results are averaged as has been done in Messing *et al.* [14]. For this data set we obtain with our method an average multi-class classification accuracy of 74%. In comparison, the authors have reported an accuracy

Action	Laptev (Harris-Laplace+HoF)	Region
AnswerPhone	19.1161%	**21.8692%**
DriveCar	80.1843%	**84.4855%**
Eat	**60.2143%**	49.6258%
FightPerson	**72.3462%**	59.2258%
GetOutCar	**25.5588%**	24.0142%
HandShake	**18.9112%**	12.2471%
HugPerson	**32.06%**	21.3516%
Kiss	47.8484%	**49.3406%**
Run	**68.8282%**	61.7558%
SitDown	**49.1482%**	40.9759%
SitUp	9.9289%	**20.7945%**
StandUp	49.0278%	**50.4238%**
Mean AP	**44.431%**	**41.3424%**

Table 2. Classification performance for *Hollywood2* data set.

of 63% based on velocity histories only and an accuracy of 67% by using latent velocity histories. Note that they also explicitly use tracking. However, our region based tracking performs better than tracking of feature points. They also report an accuracy of 89% obtained by augmenting the basic features with additional information like birth/death rate of features, and relative face information from a face detector, appearance description and colour information. These additional cues could also be used for improving the accuracy of our system.

The confusion table for this data set is given in Table 1. Our method performs well in actions that are very distinct, like *answerPhone*, *chopBanana* and *writeOnWhiteBoard* and *drinkWater*. Other actions like *eatBanana* and *dialPhone* are semantically very different, however, the performance of all these actions includes the hand being around the same position and undergoing a slight upward motion. Since the major motion is the slight upward movement of the hand, our method easily confuses the classification among these actions.

For the *Hollywood2* data set we use a one-against-all SVM classification. Each action is learned by a binary classifier. We have also experimentally compared our method to space-time interest points detected by the Harris-Laplace detector and described by HoF as done by Laptev *et al.* [12]. The *Hollywood2* data set is a significantly more challenging data set. There are videos that feature a wide range of camera motion, peripheral actions in the clips and a large variance in viewpoints and action sequences. Inherently, this data set is especially difficult for a region-tracking based approach. However, we are able to show that with the proposed approach we obtain a performance that is competitive with state of the art space time interest point (STIP) based methods. The performance of the system and its comparison with STIP features for each action is provided in Table 2. The results show that the region based approach is complementary to interest point based approach. It is also significant to note that the proposed approach is especially suited when it comes to understanding subtle actions like *Sit-Up*. The poor performance on actions like *HandShake* is mainly due to the challenging nature of the data set (for e.g. video not featuring the hands in performing the *HandShake* action).

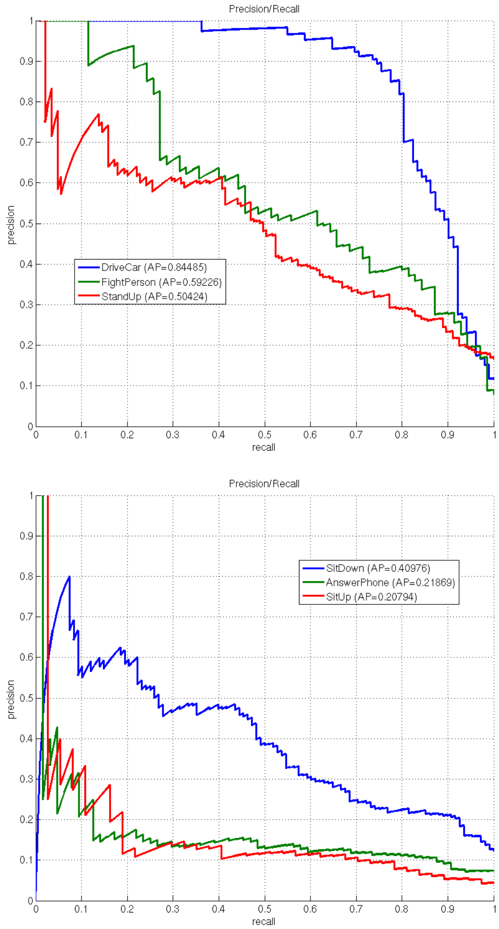


Figure 6. Representative Precision/Recall curves for *Hollywood2* data set.

6. Discussion

The proposed region descriptor has been evaluated in section 5 and compared to space-time interest point features and tracked point features. The region based features clearly show better performance than tracked point features. However, for the challenging *Hollywood2* data set, on average, STIP features perform better. Given this, it is pertinent to consider whether region based descriptors hold merit. There are several points that suggest that it is worthwhile to further explore the proposed approach: a) With region based descriptors one can obtain better spatial localization of the action. b) Obtaining the region trajectory ensures that the relevant region is covered. Hence one can use a denser extraction of features along the trajectory. c) The knowledge of the trajectory enables us to more easily discriminate peripheral features not related to the action being performed. d) Continuous feature sets can help us in explicit temporal modeling of the action using latent variables. Additionally there are implementational aspects that can be improved like better motion segmentation and optical flow components. We are also interested in exploring the joint performance of interest point and region based approaches.

7. Conclusion

In this paper we have proposed a region based action recognition system. We have shown that regions detected using motion segmentation can be described with region descriptors when they are robustly tracked. The trajectories yield stable motion flow patterns that are meaningful for understanding the action performed. Moreover, once described using contiguous region descriptors they can be matched reliably. We thus provide the basis for further exploration of local region descriptors for action recognition. We have evaluated our system on challenging state-of-the-art real-world data sets and have obtained good performance on them.

References

[1] N. Ahuja and S. Todorovic. Connected segmentation tree a joint representation of region layout and hierarchy. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[2] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 2008.

[3] C. Bibby and I. Reid. Robust real-time visual tracking using pixel-wise posteriors. In *European Conference on Computer Vision (ECCV)(2)*, pages 831–844, 2008.

[4] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 23(3):257–267, March 2001.

[5] G. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. In *IEEE Workshop on Applications of Computer Vision*, pages 174–184, 2000.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 1–8, 2005.

[7] A. A. Efros, A. C. Berg, G. P. Mori, and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 726–734, 2003.

[8] G. Farnebäck. Very high accuracy velocity estimation using orientation tensors, parametric motion, and segmentation of the motion field. In *ICCV '01: Proceedings of the Eigth IEEE International Conference on Computer Vision*, volume 1, pages 171–177, 2001.

[9] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik. Recognizing using regions. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2009.

[10] A. Kläser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 1–10, 2008.

[11] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE International Conference on Computer Vision (ICCV)(1)*, pages 432–439, 2003.

[12] I. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[13] M. Marszaek, I. Laptev, and C. Schmid. Actions in context. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2009.

[14] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2009. IEEE Computer Society.

[15] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *International Conference on Pattern Recognition (ICPR)(3)*, pages 32–36, 2004.

[16] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[17] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, page 127, sep 2009.

[18] G. Willems, J. H. Becker, T. Tuytelaars, and L. V. Gool. Exemplar-based action recognition in video. In *British Machine Vision Conference*, pages 1–11, sep 2009.

[19] G. Willems, T. Tuytelaars, and L. V. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision (ECCV)*, pages 650–663, 2008.

[20] Y. Yacoob and M. Black. Parameterized Modeling and Recognition of Activities. *Computer Vision and Image Understanding*, 73(2):232–247, 1999.

[21] L. Yefet and L. Wolf. Local trinary patterns for human action. In *ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2009. IEEE Computer Society.

[22] L. Zelnik-Manor and M.Irani. Statistical analysis of dynamic actions. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(9):1530–1535, September 2006.