Bag of Optical Flow Volumes for Image Sequence Recognition[1]

Hayko Riemenschneider
http://www.icg.tugraz.at/Members/hayko
Michael Donoser
http://www.icg.tugraz.at/Members/donoser
Horst Bischof
http://www.icg.tugraz.at/Members/bischof

Institute for Computer Graphics and Vision
Graz University of Technology
Inffeldgasse 16/II, 8010 Graz
Austria

Abstract

This paper introduces a novel 3D interest point detector and feature representation for describing image sequences. The approach considers image sequences as spatio-temporal volumes and detects Maximally Stable Volumes (MSVs) in efficiently calculated optical flow fields. This provides a set of binary optical flow volumes highlighting the dominant motions in the sequences. 3D interest points are sampled on the surface of the volumes which balance well between density and informativeness. The binary optical flow volumes are used as feature representation in a 3D shape context descriptor. A standard bag-of-words approach then allows building discriminant optical flow volume signatures for predicting class labels of previously unseen image sequences by machine learning algorithms. We evaluate the proposed method for the task of action recognition on the well-known Weizmann dataset, and show that we outperform recently proposed state-of-the-art 3D interest point detection and description methods.

1 Introduction

3D interest point detectors and descriptors for image sequences have recently been in the scope of several researchers as they constitute the basis for different computer vision applications like tracking or recognizing actions and events. The underlying idea of most approaches is to view videos as spatio-temporal volumes and therefore many proposed methods simply extend ideas successfully applied in the 2D image domain to the third dimension.

Calculation of 3D interest point descriptors consists of three different steps: definition of the underlying feature representation, detection of 3D interest points and description of the spatio-temporal volumes, which surround the interest points.

The most important part of recognizing image sequences based on 3D interest point descriptors is the choice of the underlying feature representation. Different features were used ranging from simple pixel intensity values [2], over common gradient magnitudes and orientations [8, 22] to optical flow based features as used in [4, 10]. Especially optical flow has recently proven to be a strong feature for the task of action recognition [10]. Highly accurate dense optical flow can now be calculated in real-time [26].

As outlined by [2], for interest point detection in spatio-temporal volumes, direct 3D counterparts to commonly used 2D interest point detectors are inadequate. Thus, different alternatives were proposed. The simplest approach is to sample points on a fixed grid or even randomly distributed which achieves good results if enough points are sampled. In [9] significant intensity variations in both spatial and temporal direction are detected in the style of Harris corner detection. In [2] a detector was proposed based on the response of Gabor filters applied to the spatial and the temporal domain. Recently, saliency analysis was utilized for detecting more evenly distributed 3D interest points [20].

For describing the space-time volumes around the detected interest points mainly straightforward extensions of 2D concepts were proposed. For example the well-known SIFT and the Histogram of Gradients (HoG) descriptors were directly converted to their 3D counterparts in [22] and [8] and showed impressive performance for recognizing image sequences.

Once 3D interest points are detected and the surrounding spatio-temporal volume is described in terms of fixed-size feature vectors, any standard image recognition method can be applied. The main paradigm in this field is the well-known bag-of-words model [23], where descriptors are clustered into prototypes and each image (video) is represented by its signature of the occurring prototypes. Then common machine learning algorithms like support vector machines are used to classify new videos, for example to recognize specific actions or events in the sequence.

We also follow the trend to view videos as spatio-temporal volumes and to use 3D interest point descriptors in a bag-of-words model for recognizing image sequences. Our main focus lies on introducing an interest point detection and description method based on a novel underlying feature representation. Our representation analyzes the magnitudes of the estimated optical flow fields in the sequences. Instead of directly using the flow as feature representation as done in [4, 10], we first detect connected binary volumes in the optical flow magnitude fields of the sequence. This binary representation then allows sampling 3D interest points on the surface of the volume and using the binary optical flow volumes as underlying feature representation. Please note that unlike other methods we solely exploit the optical flow as feature and do not use any appearance information in our method but nevertheless achieve state-of-the-art performance for the task of action recognition as it is shown in the experimental section. Furthermore, since all steps of our method have shown to be real-time capable by themselves, the proposed approach potentially allows real-time image sequence recognition.

The outline of the paper is as follows. Section 2 introduces our novel feature representation for 3D interest point detection and description and summarizes the bag-of-words model which is used to classify new image sequences. Section 3 provides an experimental evaluation of the proposed method for the task of action recognition. We compare to state-of-the-art methods and demonstrate improved performance despite the simplicity and high efficiency of the method.

2 Image sequence description by bag of flow volumes

Our method for recognizing image sequences consists of two steps. Section 2.1 introduces our novel optical flow based feature representation and 3D interest point detection and de-

[1]This work was supported by the Austrian Research Promotion Agency (FFG) project FIT-IT CityFit (815971/14472-GLE/ROD) and the Austrian Science Fund (FWF) under the doctoral program Confluence of Vision and Graphics W1209.

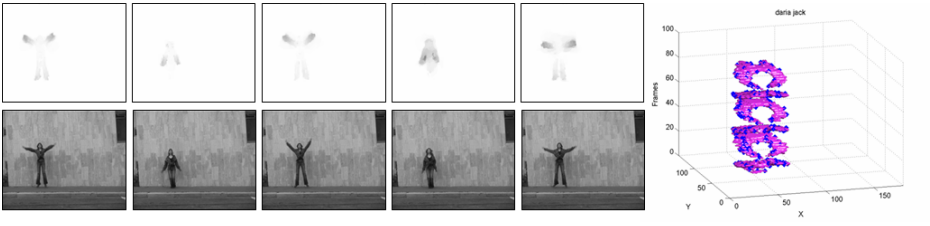

Figure 1: Illustration of 3D image sequence description process: First, the optical flow in the image sequence is calculated (top row). Second, Maximally Stable Volumes (MSVs) are detected within the flow fields and used as feature representation (on right). Third, on the surface of these stable motion volumes, 3D interest points are detected (blue dots on right) and described by a 3D shape context method.

scription method, which provides a set of fixed-size feature vectors for the detected 3D interest points in the sequence. The second part of our method follows a standard bag-of-words model which describes the image sequences in terms of optical flow volume signatures and is summarized in Section 2.2.

2.1 3D interest point detection and description

The main contribution of this paper is a novel feature representation based on analysis of the optical flow in a sequence, which is used for strong 3D interest point detection and additionally constitutes the basis for calculating a discriminant spatio-temporal descriptor.

Our method for 3D interest point detection and description mainly consists of three subsequent steps. First, we estimate the optical flow in the image sequence. Second, we apply the Maximally Stable Volume (MSV) detector to identify stable optical flow volumes, and finally we describe sampled interest points located on the volume surfaces with a 3D descriptor analyzing the local shape of the volumes.

The first step is the estimation of the optical flow within the image sequence. We apply a recently proposed TV-L_1 based variational method [26] which is one of the best performing algorithms on the well-known Middlebury dataset. For further analysis we only consider the pixel-wise magnitude fields neglecting the orientation information.

The next step of our method is to detect stable connected volumes within the calculated optical flow magnitude fields, which is done by Maximally Stable Volume (MSV) detection. MSVs were proposed by Donoser and Bischof [3] for the task of 3D segmentation. It is an extension of the Maximally Stable Extremal Region (MSER) interest region detector from Matas et al. [14] to the third dimension. MSER detection returns a set of connected regions within a gray-scale image which are defined by an extremal property of the intensity function. MSERs have properties that form their superior performance as stable local detector. MSERs are invariant to affine intensity transformations and covariant to adjacency preserving (continuous) transformations on the image domain. Furthermore, they are detected at different scales and since no smoothing is involved, both very fine and very large structures

are detected. Finally, they have shown to be the interest region detector with the lowest algorithmic complexity and therefore can be used in real-time frameworks.

Maximally Stable Volume (MSV) detection is a straight-forward extension of the MSER approach to the third dimension. Thus instead of detecting maximally stable regions in 2D, MSV detection returns the maximally stable volumes in 3D datasets. Analogue to MSERs, high stability is defined as homogeneous intensity distribution inside the volume and high intensity difference to its boundary. The detected MSVs possess the same desired properties as single image MSERs and in addition allow handling topological changes of regions in the image sequence.

The detection of MSVs within a 3D dataset is done by the same algorithm as for MSERs. It is based on interpreting the input as connected, weighted graph, where voxels are the nodes and edges are defined by for example the 6, 18 or 26 neighborhood. Then a data-structure denoted as component tree is built, which analyzes how binary threshold results of the volume change during adapting the threshold value. Each node of the component tree contains a volume and the tree structure allows calculating a stability value for every node analyzing how much the size of the volume changes while moving the component tree upwards. The most stable volumes, i. e. the nodes with the highest stability values, are returned as detection result. The calculation of the component tree and the analysis of the most stable volumes can be done in an efficient manner, for example Nistér and Stewénius [17] recently showed a linear time algorithm for detection of MSERs, which can be easily extended to the third dimension.

We apply the MSV detector on the calculated optical flow magnitude fields interpreting the image sequence as a 3D dataset. The final output is a set of stable connected flow volumes for each image sequence as it is illustrated in Figure 2 for videos of the Weizmann action recognition dataset [1].

As next step we randomly sample 3D interest points located at the surface of the detected volumes, therefore focusing on areas where the optical flow magnitude changes in the sequence. Each interest point is described by a 3D descriptor, where any of the versatile methods available can be used. We use a 3D shape context descriptor, which allows describing the local shape of the binary volumes. The 3D shape context as proposed by [8] is a straight-forward extension of the common 2D shape context by log-polar binning of the surface points in the spatio-temporal neighborhood.

It is important to note, that in contrast to almost all other 3D interest point description methods we do not use the appearance or gray scale images, we instead propose to use the detected binary Maximally Stable Volumes (MSV) as underlying representation. Therefore, we mainly analyze the local surface shape of the optical flow volumes. As it is shown in the experiments in Section 3, using the binary optical flow volumes as representation significantly improves recognition results.

2.2 Bag of optical flow volumes

We now have a set of fixed-size 3D interest point description vectors for every input video sequence. This allows applying a standard bag-of-words model for recognizing and distinguishing different image sequences. The bag-of-words model for recognizing videos has been used several times before as for example by [5, 10, 21].

The underlying concept of a bag-of-words model is to represent images (videos) by counting the number of occurrences of descriptor prototypes, so-called visual words. Visual words are identified by clustering the set of all training descriptors to identify their shared

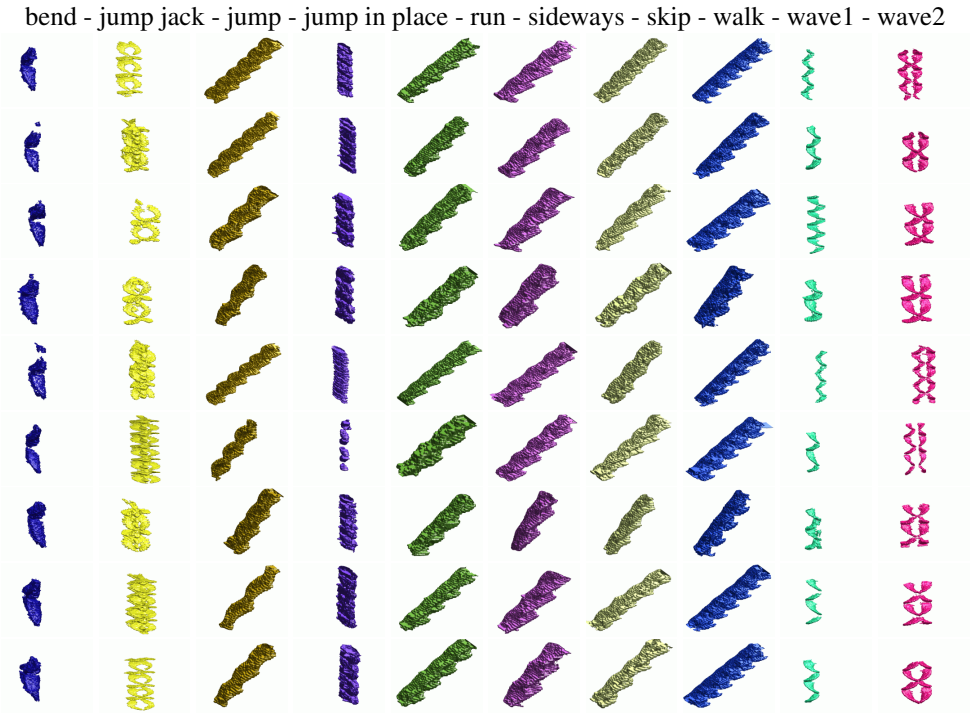

Figure 2: Illustration of the Maximally Stable Volumes (MSV) extracted from the optical flow fields of the action sequences from the Weizmann dataset. Each row contains the stable flow volumes calculated from the image sequences of the nine subjects. The ten columns show the corresponding ten different actions. Note how each flow volume has a unique volume surface, even the minor differences between sideways galloping and skipping are visible when looking at the motion of the legs and feet.

properties. Histograms of the visual word occurrences define the basic building blocks for comprising and identifying images (videos). For efficiency reasons we use the properties of a hierarchical k-means clustering approach proposed by Nistér and Stewénius [16]. The idea is to build a tree structure by repeated clustering of the descriptors to get various levels of abstraction. This property is used to quickly discard a large number of descriptors when searching for the best match. At each level only the cluster with the most similar visual word is further considered. This creates a significant speedup and has been shown to work for one million images at a query time of less than one second [16].

In this work we cluster the description vectors of the detected 3D interest points into meaningful visual words representing local optical flow volume prototypes. By counting the number of visual word occurrences in the image sequence, we build a discriminative signature which can be used for classifying previously unseen sequences. Figure 3 illustrates the coherency of the visual words obtained for the Weizmann dataset. It shows a set of selected clusters and their associated optical flow volumes surrounding the corresponding interest points. The low intra-class variance demonstrates the power of our approach as similar flow

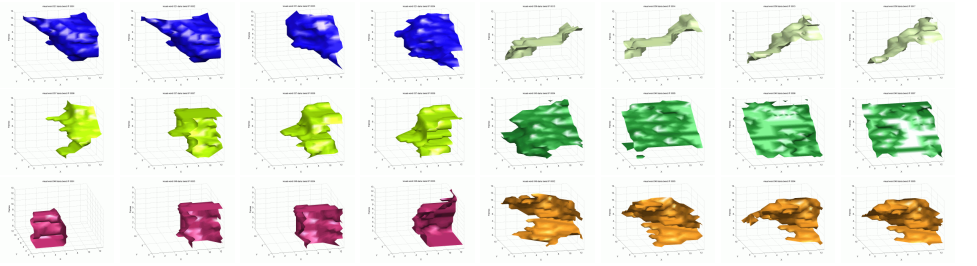

Figure 3: Illustration of obtained visual word clusters. Each of the colored clusters (here a subset of the overall 243 clusters) shows examples of parts of the 3D optical flow volumes as clustered together by the hierarchical k-means method. These visual words are used for identifying the characteristic motion parts of each image sequence.

volumes are clustered together to identify characteristic motions in the sequences.

In the final stage a machine learning algorithm is used to train and classify previously unseen visual word signatures. Usually, a support vector machine is used for this step which is a binary classifier. Therefore, to be able to tackle multi-class problems these approaches have to be extended by common techniques like 1-vs.-all, 1-vs.-1 or error correcting output codes (ECOC). We have chosen the randomized ferns method proposed by Özuysal et al. [19], because it is implicitly multi-class capable and has shown to be impressively efficient. The random fern is a semi-naive Bayesian classifier and an adaptation of the random forest concept introduced by Lepetit et al. [12]. The underlying idea is to build a large number of random decision trees, where each node represents a decision that narrows down the choices to a final decision outcome. Random ferns model the conditional probabilities derived from a large number of binary decisions based on simple single feature comparisons. Similar to random forests they can handle high dimensional datasets and do not suffer from overfitting. In our method we use decision stumps of single visual word histogram bins as the internal node tests. The number of leaf nodes and the overall number of ferns is fixed for all experiments and parameters are given in Section 3.1.

Please note that all five required steps of our image sequence recognition method all for themselves have shown to be real-time capable: optical flow detection in [26], MSV detection in linear time in [17], 3D shape context in [8], hierarchical k-means clustering in [16] and random ferns in [18]. Thus, considering that optical flow volumes can be calculated incrementally by only considering the currently required number of frames (defined by the temporal scale of the 3D shape context descriptor), a real-time image sequence recognition framework could possibly be implemented.

3 Experiments

To evaluate the performance of our proposed 3D interest points, we applied it for the task of action recognition. We used the well-known Weizmann dataset [1] and use an experimental setup analogue to [8, 22]. The dataset consists of ten different types of actions:

IP detection / Feat+Descriptor	gray+3D SIFT	flow+3D SIFT	binary MSV+SC
random sampling	87.78%	90.44%	93.09%
sampling on volume surface	90.22%	93.11%	**96.67%**

Table 1: Comparison of the action recognition performance on the Weizmann dataset for different combinations of interest point detection and feature representation + description methods. We evaluated the performance using the pure gray-scale appearance, the magnitude of the optical flow (both described with 3D SIFT [22]) and our binary flow volume (described by a 3D shape context [6]) at the same 3D interest point locations. The 3D locations are detected by random sampling or selected on the surface of the optical flow volumes.

bending, jumping jack, jumping, jump in place, running, side jumping, skipping, walking, one-hand and two-hand waving. There exist videos for each action by nine subjects. Testing is performed in a leave-one-out-fashion on a per person basis, i. e. training is done on eight subjects and testing on the unused subject and all its videos.

3.1 Experimental procedure and parameters

The procedure is divided in the following six steps: First, the optical flow is calculated in each video file. Second, within the optical flow magnitude fields the Maximally Stable Volumes (MSVs) are detected with a minimum size of 200 voxels and under a stability $\Delta = 5$. These volumes are used as feature representation. Third, interest point selection is done in two variants. For comparison with [22] we either sample random points around the binary volumes or we select random points located on the surface of the volumes. The number of points is equal in both cases and depends on the complexity of the surface as we sample around 5 percent of the points on the surface volume. This results in 350 feature points on average per sequence or six features points per frame. Fourth, the interest points are described using two different descriptor variants: the 3D SIFT description method of [22] and a 3D implementation of shape context [6]. The dimension of the descriptors are chosen to be equal in all directions and amount to spatial and temporal support of $d_s = d_t = 12$. The underlying feature representation is either our binary flow volume or, again for comparison reasons, the gray-scale image data or the estimated optical flow magnitudes. Fifth, all descriptors for all videos are clustered with a hierarchical k-means into 243 visual words with k = 3. The signatures for these visual words are used to classify the test videos. Sixth, the final classification decision is done using random ferns, which are trained with 150 trees each with 1024 leaf nodes.

The entire runtime performance of the recognition for an unseen image sequence is about 1.4 seconds per frame in our non-optimized Matlab implementation. The slowest components here are the calculation of the MSVs with 480 ms and the 3D shape context description with 760 ms on a single 3 GHz core.

3.2 Results

First, average recognition results are given in Table 1 comparing different combinations of interest point detection and feature representation + descriptor variants. As can be seen sam-

ST features [13]	68.4%
Shape Context+Grad+PCA [15]	72.8%
Spin images [13]	74.2%
3D SIFT [22]	82.6%
Klaeser [8]	84.3%
Our method	**96.7%**

Table 2: Average recognition rates for the Weizmann actions for state-of-the-art methods using 3D interest point descriptors. Our method is able to boost the recognition performance on the entire video when comparing to related work using a bag-of-words model. Please note, more complex methods like [1, 7, 21, 24, 25] already achieve up to 100% on the Weizmann dataset, however, sometimes only using nine of the ten actions. These results prove that simple optical flow volumes are a competitive feature representation for the task of action recognition.

pling on the surface of the volume always outperforms random sampling around the visual motion areas. Using the interest points located on the surface of the stable flow volumes improves the results by 3%. This clearly shows the benefit of analyzing stable flow volumes for detecting 3D interest points. The main reason for the improved performance mainly seems to be that uniformly sampling on the volumes achieves a good balance between interest point density and informativeness.

In the columns of Table 1 we compare different feature representation and descriptor combinations. We evaluated three different variants. First, gray-scale image data is described with a 3D SIFT descriptor. Second, the optical flow magnitude fields are also described by 3D SIFT. Finally, our binary volumes are described using a 3D shape context. The best result is achieved by our binary stable flow volume and 3D shape context description. This demonstrates the power of using simple binary flow volumes as feature representation and for interest point detection. The optical flow volume surfaces alone carry enough information to correctly classify image sequences. Please note that in contrast to other bag-of-words approaches [2, 9, 13] we do not require any high-level features or segmentations of the moving subjects.

The second part of the results shows a comparison to state-of-the-art work in Table 2. In comparison to directly related work using 3D interest point descriptors in bag-of-words models for classifying image sequences, we can improve the results by more than 10% to 96.7%. To the best of our knowledge, this is the highest reported score for action recognition on the Weizmann dataset using the simple bag-of-words model of 3D interest point descriptors. Please note that much more complex methods which for example combine multiple features, analyze spatial relationships or even exploit provided binary segmentations in every frame like [1, 7, 21, 24, 25] already achieved up to 100% on this dataset. However, some of these scores were only achieved for the original nine actions, where the most difficult action (skip) is not included.

Figure 4 shows a confusion matrix, which illustrates our excellent results on the Weizmann action dataset. As can be seen there are only three confusions between actions. One exists between one-hand and two-hand waving, which results from the frequency of the otherwise very similar flow volume words. We are able to almost correctly classify the problem-

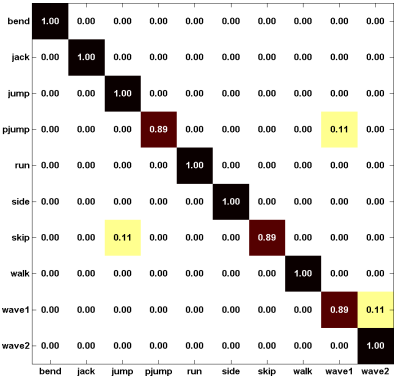

Figure 4: Confusion map of our method for the ten Weizmann actions: bending, jumping jack, jumping, jump on position, running, side jumping, skipping, walking, one-hand and two-hand waving. We are able to classify 96.7% of the image sequences correctly while only confusing the difficult skipping action once. The confusion between the waving actions is due to the simple histogram of visual words, which amounts to twice as high frequencies for the two-hand motion. The third confusion exists between the visual similarities (see Figure 2) of the jumping-in-place to the waving. However, please note that these confusions could be easily resolved if spatial relations were added.

atic action of skipping, and just one confusion with the sideways galloping is recorded. The final yet most interesting confusion shows between the jumping-in-place action and the one-hand waving. This may be attributed again to similar visual flow words, when comparing the flow volumes in columns four and nine shown in Figure 2.

4 Conclusion and outlook

In this paper we have introduced a novel 3D interest point detector and feature representation for describing image sequences. Each image sequence is analyzed as spatio-temporal volume and Maximally Stable Volumes (MSVs) are detected in efficiently calculated optical flow magnitude fields. The resulting binary flow volumes highlight the dominant motions in the sequences. The main contribution of this paper is that we demonstrate that these simple binary volumes are sufficient to recognize image sequences as they serve as underlying feature representation and as basis for detecting 3D interest points. By analyzing signatures of the optical flow volume prototypes in a bag-of-words model, we can classify previously unseen sequences. We show results for the task of action recognition on the well-known Weizmann dataset. The proposed method outperforms recently proposed state-of-the-art 3D interest point detection and description approaches and based on the simple bag-of-words model achieves an excellent overall recognition performance of 96.7%. Future work will focus on combining our simple binary optical flow features with appearance-based features and on integrating spatial distributions of the descriptors for further improving recognition performance.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1395–1402, 2005.

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 65–72, 2005.

[3] M. Donoser and H. Bischof. 3D Segmentation by Maximally Stable Volumes (MSVs). In *Proceedings of International Conference on Pattern Recognition (ICPR)*, pages 63–66, 2006.

[4] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 726–733, 2003.

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29 (12):2247–2253, 2007.

[6] M. Grundmann, F. Meier, and I. Essa. 3D Shape Context and Distance Transform for Action Recognition. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2008.

[7] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A Biologically Inspired System for Action Recognition. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2007.

[8] A. Klaeser, M. Marszałek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In *Proceedings of British Machine Vision Conference (BMVC)*, 2008.

[9] I. Laptev and T. Lindeberg. Space-time Interest Points. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 432–439, 2003.

[10] I. Laptev and T. Lindeberg. Local Descriptors for Spatio-Temporal Recognition. In *International Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[12] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 775–781, 2005.

[13] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[14] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proceedings of British Machine Vision Conference (BMVC)*, pages 384–393, 2002.

[15] J. Niebles and L. Fei-Fei. A Hierarchical Model of Shape and Appearance for Human Action Classification. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[16] D. Nistér and H. Stewénius. Scalable Recognition with a Vocabulary Tree. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, 2006.

[17] D. Nistér and H. Stewénius. Linear Time Maximally Stable Extremal Regions. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 183–196, 2008.

[18] M. Özuysal, P. Fua, and V. Lepetit. Fast Keypoint Recognition in Ten Lines of Code. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[19] M. Özuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.

[20] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[21] K. Schindler and L. van Gool. Action snippets: How many frames does human action recognition require? In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[22] P. Scovanner, S. Ali, and M. Shah. A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition. *ACM Multimedia*, 2007.

[23] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003.

[24] L. Wang and D. Suter. Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[25] D. Weinland and E. Boyer. Action Recognition using Exemplar-based Embedding. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[26] C. Zach, T. Pock, and H. Bischof. A Duality Based Approach for Realtime TV-L1 Optical Flow. In *Proceedings of Symposium of the German Association for Pattern Recognition (DAGM)*, pages 214–223, 2007.