Bag of Optical Flow Volumes for Image Sequence Recognition

Hayko Riemenschneider[1]
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
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 (MSV) in efficiently calculated optical flow fields. This provides a set of binary optical flow volumes highlighting the dominant motions in the sequences. Embedded in a simple bag-of-words model, we achieve excellent results on the Weizmann dataset outperforming recent 3D interest point detection and description methods.

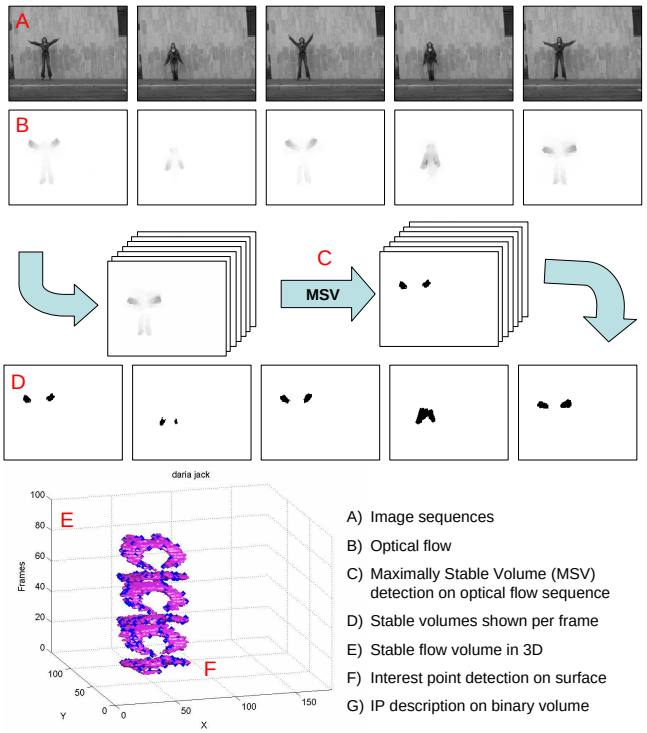

Figure 1: Illustration of image sequence description: First, optical flow is calculated in the image sequence. Second, Maximally Stable Volumes (MSV) are detected within the magnitude flow fields and used as feature representation. Third, on the surface of these stable motion volumes 3D interest points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

Stable optical flow volume: Our approach analyzes the magnitudes of the estimated optical flow fields in the sequences. Instead of directly using the flow as feature representation as done in [2], we first detect connected binary volumes in the optical flow magnitude fields of the sequence. For this we apply Maximally Stable Volume (MSV) detection [1] which is an extension of Maximally Stable Extremal Region (MSER) interest region detector to the third dimension. These binary stable optical flow volumes are used as underlying feature representation.

Interest point detection and description: 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. The first part of our method introduces an interest point detection and description method based on the novel underlying feature representation of binary stable flow volumes. The MSV representation allows sampling of 3D interest points on the surface of the optical flow volume as well as description based on a 3D shape context method of the local binary volume. The second part follows a standard bag-of-words model which describes the image sequences in terms of optical flow volume signatures.

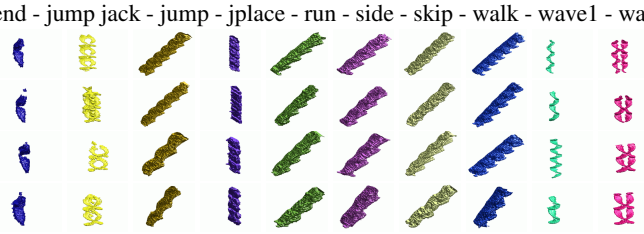

Figure 2: Illustration of the Maximally Stable Volumes (MSV) extracted from the optical flow fields of the action sequences from the Weizmann dataset. Note how each MSV has a unique volume surface, even the minor differences between sideways galloping and skipping are visible.

Pure binary volume: 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.

Experiments: To evaluate the performance of our proposed 3D interest points, we applied it for the task of action recognition. The stable binary flow volumes are used as feature representation and we either sample random points or we select points on the surface of the volumes. Different combinations of interest point detection and feature representation and descriptor methods are compared. See full paper for details.

Shape Context [4]	3D SIFT [5]	Klaeser [3]	Our method
72.8%	82.6%	84.3%	**96.7%**

Table 1: Average recognition rates for the Weizmann dataset. Our method is able to boost the performance on the entire video when compared with related work using 3D interest points in a bag-of-words model. These results prove that simple optical flow volumes are a competitive feature representation for the task of image sequence recognition.

Conclusion: The main contribution of this paper[1] is a novel feature representation based on MSV analysis of the optical flow in an image sequence. The resulting stable binary flow volumes are used for strong 3D interest point detection and additionally calculating a discriminant spatio-temporal descriptor. We demonstrate that interest points on the surface of these simple binary volumes are sufficient to recognize image sequences in a bag-of-words model. In Table 1 we show an excellent overall recognition performance of 96.7% for the task of action recognition on the well-known Weizmann action dataset.

[1] M. Donoser and H. Bischof. 3D Segmentation by Maximally Stable Volumes (MSVs). In *ICPR*, pages 63–66, 2006.

[2] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing Action at a Distance. In *ICCV*, pages 726–733, 2003.

[3] A. Klaeser, M. Marszałek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In *BMVC*, 2008.

[4] J. Niebles and L. Fei-Fei. A Hierarchical Model of Shape and Appearance for Human Action Classification. In *CVPR*, 2007.

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

[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.