Int J Comput Vis (2013) 103:60–79
DOI 10.1007/s11263-012-0594-8

Dense Trajectories and Motion Boundary Descriptors for Action Recognition

Heng Wang · Alexander Kläser · Cordelia Schmid · Cheng-Lin Liu

Received: 27 December 2011 / Accepted: 19 October 2012 / Published online: 6 March 2013

Abstract This paper introduces a video representation based on dense trajectories and motion boundary descriptors. Trajectories capture the local motion information of the video. A dense representation guarantees a good coverage of foreground motion as well as of the surrounding context. A state-of-the-art optical flow algorithm enables a robust and efficient extraction of dense trajectories. As descriptors we extract features aligned with the trajectories to characterize shape (point coordinates), appearance (histograms of oriented gradients) and motion (histograms of optical flow). Additionally, we introduce a descriptor based on motion boundary histograms (MBH) which rely on differential optical flow. The MBH descriptor shows to consistently outperform other state-of-the-art descriptors, in particular on real-world videos that contain a significant amount of camera motion. We evaluate our video representation in the context of action classification on nine datasets, namely KTH, YouTube, Hollywood2, UCF sports, IXMAS, UIUC, Olympic Sports, UCF50 and HMDB51. On all datasets our approach outperforms current state-of-the-art results.

Keywords Action recognition · Dense trajectories · Motion boundary histograms

H. Wang (✉)· C.-L. Liu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: hengwang00@gmail.com; hwang@nlpr.ia.ac.cn

C.-L. Liu
e-mail: liucl@nlpr.ia.ac.cn

A. Kläser · C. Schmid
LEAR Team, INRIA Grenoble Rhône-Alpes, 655 Avenue de l'Europe, 38330Montbonnot, France
e-mail: alexander.klaser@inria.fr

C. Schmid
e-mail: cordelia.schmid@inria.fr

1 Introduction

Local space-time features are a successful representation for action recognition. Laptev (2005) has introduced space-time interest points by extending the Harris detector to video. Other detection approaches are based on Gabor filters (Bregonzio et al. 2009; Dollár et al. 2005) and on the determinant of the spatio-temporal Hessian matrix (Willems et al. 2008). Feature descriptors range from higher order derivatives (local jets), gradient information, optical flow, and brightness information (Dollár et al. 2005; Laptev et al. 2008; Schüldt et al. 2004) to spatio-temporal extensions of image descriptors, such as 3D-SIFT (Scovanner et al. 2007), HOG3D (Kläser et al. 2008), extended SURF (Willems et al. 2008), and Local Trinary Patterns (Yeffet and Wolf 2009).

However, the 2D space domain and 1D time domain in videos show different characteristics. It is, therefore, more intuitive to handle them in a different manner than to detect interest points in a joint 3D space. Tracking interest points through video sequences is a straightforward choice. Some recent methods (Matikainen et al. 2009; Messing et al. 2009; Sun et al. 2009, 2010) show good results for action recognition by leveraging the motion information of trajectories. To obtain feature trajectories, either tracking techniques based on the KLT tracker (Lucas and Kanade 1981) are used (Matikainen et al. 2009; Messing et al. 2009), or SIFT descriptors between consecutive frames are matched (Sun et al. 2009). Recently, Sun et al. (2010) combined both approaches and added random trajectories in low density regions of both trackers in order to increase density.

Dense sampling has shown to improve results over sparse interest points for image classification (Fei-Fei and Perona 2005; Nowak et al. 2006). The same is observed for action recognition in a recent evaluation by Wang et al. (2009), where dense sampling at regular positions in space and time