Combined trajectories for action recognition based on saliency detection and motion boundary

Xiaofang Wang, Chun Qi, Fei Lin

PII: S0923-5965(17)30091-7
DOI: http://dx.doi.org/10.1016/j.image.2017.05.007
Reference: IMAGE 15226

To appear in: *Signal Processing: Image Communication*

Received date : 10 November 2016
Revised date : 11 May 2017
Accepted date : 11 May 2017

Please cite this article as: X. Wang, et al., Combined trajectories for action recognition based on saliency detection and motion boundary, *Signal Processing: Image Communication* (2017), http://dx.doi.org/10.1016/j.image.2017.05.007

Combined trajectories for action recognition based on saliency detection and motion boundary

Xiaofang Wang[a,b], Chun Qi[a,*], Fei Lin[b]

[a]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[b]School of Electrical Engineering and Automation, Qilu University of Technology, Jinan, Shandong 250353, China

Abstract

To exploit the trajectories from different areas of a video in an effective way to represent action, this paper proposes to extract the trajectories of action-related areas, the trajectories of action-related motion boundaries and the dense trajectories independently, and then concatenate the representations of them to obtain the final representation of the video. The key to extract the former two sets of trajectories is to detect the action-related areas in each frame at first. We fulfill this task by applying sparse representation to the motion of the subvideo centered at current frame on patch level. To this end, we spatially divide the subvideo into patches. For each patch, we learn a weighted sparse representation of its motion vector using the dictionary constructed by the motion vectors of all the rest patches, and then use the reconstruction error to measure patch saliency. Based on the saliency of all patches in a frame, a saliency map is obtained to indicate the action-related areas, which on one hand is incorporated into dense tracking to extract the trajectories of action-related areas, and on the other hand is used as a mask to filter out the background motion boundaries so that the action-related motion boundary trajectories are derived. The experiments on four benchmark datasets, namely, Hollywood2, YouTube, HMDB51 and UCF101, demonstrate the effectiveness of our method.

Keywords: Action recognition, Trajectory-based method, Sparse representation, Saliency detection, Motion boundary

1. Introduction

Automatically recognizing actions in videos is involved in many applications of computer vision, such as human-computer interaction, content-based video retrieval and sports video analysis. In past decades, researchers have made great efforts to propose various methods to improve recognition performance. So far, it is not a great challenge to recognize the actions performed in controlled environments. However, for the videos captured in realistic scenes, e.g., the movie videos and sports videos, the performance of action recognition still suffers from the difficulties in extracting discriminative features due to camera movement, viewpoint change, occlusion, etc.

Recent research in action recognition demonstrates that dense trajectories are very successful in characterizing actions in realistic videos. Since Wang et al. [1, 2] extract the trajectories of densely sampled points across frames to represent actions and significantly improve the sparse key point trajectories [3–5] on several realistic action datasets, dense trajectories are frequently used as the baseline features for action recognition. For example, dense trajectory features are used in [6] as the raw motion features to determine good practices in VLAD-based video encoding; [7] proposes a new global representation, Multi-View Super Vector (MVSV), based on dense trajectories for action recognition. Despite the success, the performance of the conventional dense trajectory method is still limited due the uniform dense tracking, in which the points for tracking are densely sampled uniformly over all regions in each frame. For a video with dynamic background, dense trajectories are produced from both the action-related areas and background, see the second row of Figure 1. Background trajectories are mainly generated by the action-irrelevant motion, especially camera movement, which makes them far less informative than the trajectories in action-related areas. However the conventional dense trajectory method treats these two sets of trajectories equally in computing action representations. Obviously, this is not beneficial for improving recognition performance.

To improve the dense trajectory method, some research works [8, 9] aim to rectify the overall motion of videos to obtain the action-related trajectories. In [8], camera motion is canceled out from optical flow by estimating a homography with RANSAC, and based on the warped optical flow, the background trajectories are removed and motion-related descriptors are computed. In [9], Jain et al. separate affine flow from optical flow through a 2D affine motion model and exploit the compensated flow for action-related trajectory extraction and descriptor computation. Another improved technique extracts action-related trajectories by virtue of salient (or action-related) region detection. In [10], Vig et al. employ saliency-mapping algorithms to find informative regions and extract trajectories corresponding to these regions. In [11], Wang et al. propose to extract saliency-based dense trajectories based on the saliency maps computed by implementing low-rank matrix decomposition on motion information on a subvideo basis. All these

*Corresponding author Tel.: +86 29 82675240
Email addresses: wxf2012@stu.xjtu.edu.cn (Xiaofang Wang), qichun@mail.xjtu.edu.cn (Chun Qi), linfei@qlu.edu.cn (Fei Lin)