Egocentric Gesture Recognition

Anonymous CVPR submission

Paper ID ****

Abstract

We present a novel method for monocular hand gesture recognition in ego-vision scenarios that deals with static and dynamic gestures and can achieve high accuracy results using a few positive samples. Specifically, we use and extend the dense trajectories approach that has been successfully introduced for action recognition. Dense features are extracted around regions selected by a new hand segmentation technique that integrates superpixel classification, temporal and spatial coherence. We extensively test our gesture recognition and segmentation algorithms on public datasets and propose a new dataset shot with a wearable camera.

1. Introduction

Ego-centric vision is a paradigm that joins in the same loop humans and wearable devices to augment the subject vision capabilities by automatically processing videos captured with a first-person camera. We are interested in investigating the usage of ego-vision algorithms and devices to enhance new human-machine interfaces that could integrate information from the local environment with web and social media. These interfaces could help users to generate and share content in real-time, and could offer a customized experience, more suited for the users specific cognitive needs and interests. For instance, ego-vision wearable systems could help understand what visitors of a museum are observing or doing, and determine their degree of interest, collecting data to enhance and customize visitors experience.

Moreover, the recent growth of computational capability of embedded devices has made possible to exploit wearable and low-power devices as target platforms for egocentric real-time applications. For this reason, applications and algorithms designed for ego-vision must be suited for portable input and elaboration devices, that often present a more constrained scenario, with different power needs and performance capabilities.

2. Proposed Method

Gesture recognition systems should recognize both static and dynamic hand movements. Therefore, we propose to describe each gesture as a collection of dense trajectories extracted around hand regions. Feature points are sampled inside and around the users hands and tracked during the gesture; then several descriptors are computed inside a spatio-temporal volume aligned with each trajectory, in order to capture its shape, appearance and movement at each frame. These descriptors are coded, using the Bag of Words approach and power normalization, in order to obtain the final feature vectors, which are then classified using a linear SVM classifier.

2.1. Camera motion removal

To describe shape, appearance and movement of each trajectory we use the Trajectory descriptor, histograms of oriented gradients, of optical flow, and motion boundary histograms. The Trajectory descriptor captures trajectory shape, HOG are based on the orientation of image gradient and encode the static appearance of the region surrounding the trajectory, HOF and MBH are based on optical flow and capture motion information. In order to remove camera motion, the homography between two consecutive frames is estimated running the RANSAC algorithm on densely sampled features points. SURF features and sample motion vector are extracted from the optical flow to get dense matches between frames. However, in first-person camera views hands movement is not consistent with camera motion and this generates wrong matches between the two frames. For this reason we introduce a segmentation mask that disregards feature matches belonging to hands. In fact, without the hand seg- mentation mask, many feature points from the users hands would become inliers, degrading the homography estima- tion. As a consequence, the trajectories extracted from the video would be incorrect. Instead, computing an homogra- phy using feature points from non-hand regions allows us remove all the camera movements.

2.2. Type-style and fonts

Gesture Description Having removed camera motion between two adjacent frames, trajectories can be extracted. The second frame is warped with the estimated homography, the optical flow between the first and the second frame is recomputed, and then feature points around the hands of the user are sampled and tracked following what [18] does for human action recog- nition. Feature points are densely sampled at several spatial scales and tracked using median filtering in a dense optical flow field. In contrast to [18], trajectories are restricted to lie inside and around the users hands: at each frame the hand mask is dilated, and all the feature points outside the computed mask are discarded. Then, the spatio-temporal volume aligned with each trajectory is considered, and Trajectory descriptor, HOG, HOF and MBH are computed around it. While HOF and MBH are averaged on five consecutive frames, a single HOG descriptor is computed for each frame. In this way we can better describe how the hand pose changes in time. After this step, we get a variable number of trajectories for each gesture. In order to obtain a fixed size descriptor, the Bag of Words approach is exploited: we train four separate codebooks, one for each descriptor. Each codebook contains 500 visual words and is obtained running the k-means algorithm in the feature space.

2.3. Hand Segmentation

The proposed gesture recognition approach uses a hand segmentation mask to distinguish between camera and hand motions, and to prune away all the trajectories that do not belong to the user hand. In this way, our descriptor captures hands movement and shape as if the camera was fixed, and disregards the noise coming from other moving regions that could be in the scene. For computing hand segmentation masks, at each frame we extract superpixels using the SLIC algorithm [3], that performs a k-means-based local clustering of pixels in a 5- dimensional space, where color and pixel coordinates are used. Superpixels are represented with several features: his- tograms in the HSV and LAB color spaces (that have been proved to be good features for skin representation [8]), Ga- bor filters and a simple histogram of gradients, to discrimi- nate between objects with a similar color distribution. In order to deal with different illumination conditions we also train a collection of Random Forest classifiers indexed by a global HSV histogram, instead of using a single classi- fier. Hence, training images are distributed among the clas- sifiers by a k-means clustering on the feature space. At test time, the predictions from the five nearest classifier are av- eraged to make the final prediction. Furthermore, semantic coherence in time and space is taken into account. Since past frames should affect the pre- diction for the current frame, a smoothing filter is applied, so that the prediction for each frame is replaced with a combination of the classifier results from past frames. Then, to remove small and isolated pixel groups and also to aggregate bigger connected pixel groups, the GrabCut algorithm is applied to exploit spatial consistency.

Number of Gestures	Accuracy
2	90.6
3	85.1
4	78.2

Table 1. Results. Without Segmentation.

Number of Gestures	Accuracy
2	94.6
3	88.1
4	83

Table 2. Results. With Segmentation.

3. Results

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```