Gesture Recognition in EgoCentric Videos using Dense Trajectories and Hand Segmentation

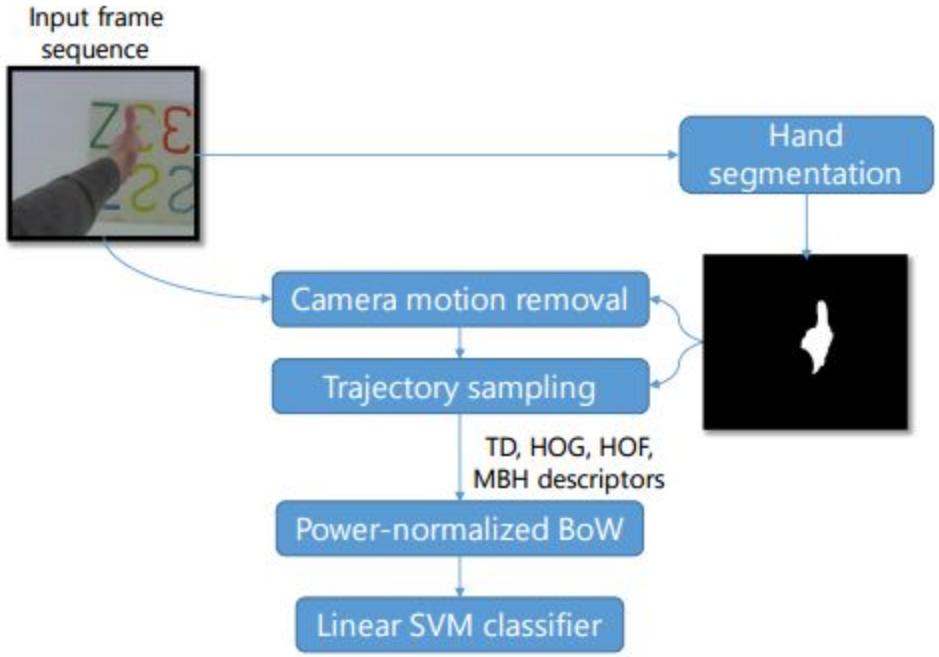
A novel method for monocular hand gesture recognition in ego-vision scenarios that deals with static and dynamic gestures and can achieve high accuracy results using a few positive samples. Specifically, we use and extend the dense trajectories approach that has been successfully introduced for action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

(a) *Dislike* gesture

(b) *Point* gesture

Different types of gestures as above were collected such as like,dislike,point,picture etc.and trained classifier.

The main idea is to classify different types of gestures which are made using hand.We use hand segmentation using a mask and extract dense trajectories and extract features around the hand and track the video and get the optical flow through it.We extract several features and concatenate them and form and big feature and power normalise it.We then run a svm classifier on it to classify the features.



Camera motion removal:

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.

Gesture Description:

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.Feature points are densely sampled at several spatial scales and tracked using median filtering in a dense optical flow field.

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.

$$f(h_i) = \mathrm{sign}(h_i) \cdot |h_i|^{\frac{1}{2}}$$

The final feature vector is then obtained by the concatenation of its four power-normalized histograms. Eventually, gestures are recognized using a linear SVM 1-vs-1 classifier.

Hand Segmentation:

For computing hand segmentation masks, at each frame we extract superpixels using the SLIC algorithm and perform 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: histograms in the HSV and LAB color spaces (that have been proved to be good features for skin representation).Gabor filters and a simple histogram of gradients, to discriminate between objects with a similar color distribution.