Surrounding Scene Appearance as Contextual Features for Action Recognition

Fabian Caba Heilbron[1,2], Victor Escorcia[1,2], Juan Carlos Niebles[2], Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[2]Universidad del Norte (UNINORTE), Colombia

Abstract. This notebook paper describes the submission of the KAUST-UNINORTE team to the challenge THUMOS 2014. Our system extracts a new set of visual cues that represent the *context* of an action as described in [1]. Using dense point trajectories, our approach separates and describes the foreground motion from the background and represents the appearance of the extracted static background that interestingly is shown to be discriminative for certain action classes. We employ a Fisher Vector to encode each descriptor separately. Finally, we concatenate all computed descriptors and learn a linear SVM classifier to predict the action labels.

Keywords: Action recognition, Fisher vectors, Contextual features

1 Pipeline

The methodology in this paper follows the conventional action recognition pipeline. Given a set of labelled videos, a set of features is extracted from each video, represented using visual descriptors, and combined into a single video descriptor, which is used to train a multi-class classifier.

We use dense point trajectories (short tracks of a densely sampled set of pixels in a video [4]) as our primitive features. By estimating frame-to-frame camera motion with a fundamental matrix, we separate foreground trajectories corresponding to the action from background ones. Each type of trajectory is represented using a different descriptor. Foreground trajectories are represented using conventional visual properties (e.g. MBH, HOF, HOG), while the surrounding scene appearance is described using SIFT. Foreground and background trajectories are then encoded separately using the BoF framework as illustrated in Figure 1.

Exploiting the well performance of the camera movement compensation, we can easily detect trajectory points associated with the background, which tend to present a small displacement over the length of the trajectory. Taking advantage of this, we threshold the trajectory displacement to obtain a foreground-background separation as follow:

$$D = \sum_{j=t}^{t+L-1} \left((x_{t+1} - x_t)^2, (y_{t+1} - y_t)^2 \right). \quad (1)$$

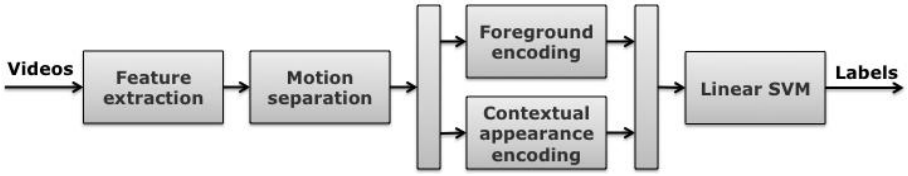

Fig. 1. Given a video sequence, a set of dense points trajectories are extracted. Then, a Fundamental Matrix is used for both applying a camera compensation and separating foreground/background trajectories. Each type of trajectories are encoded by different type of descriptors. Specifically, surrounding scene appearance is explicitly computed on background trajectories and traditional foreground descriptors (e.g. MBH, HOF, HOG) are also aggregated in the video description. Finally, this set of descriptors are encoded separately using the BoF framework using the fisher vector encoding.

Trajectory points are associated with the background if $D <= \alpha$. Otherwise, those trajectory points are labeled as foreground. Empirically, we set this threshold value to $\alpha = 3$ pixels.

In practice, we calculate **foreground descriptors** that consist of HOG, HOF, and MBH computed over improved trajectories as in [4]. Our surrounding scene appearance is encoded using SIFT descriptors [2] around trajectory points associated with the background. We detect SIFT keypoints in a dense manner. Then, we filter out those that fall within the union of foreground trajectories.

We generate the visual codebook using a **Gaussian Mixture Model (GMM)**, capturing probability distributions over feature space. We encode each descriptor applying the recently introduced Fisher vectors approach as in [3]. Finally we concatenate each fisher vector to build a linear SVM classifier to predict action labels for each video.

2 Classification task submission

To train the models used in the THUMOS challenge classification task, we use both the training subset (UCF101) and the THUMOS validation subset. Due extensive computations required to extract features on validation set, we apply mean-shift clustering over the temporal duration of the training videos in order to get representative windows length over all the action categories. Then, we randomly extract clips of different windows length for each video from the validation set. The same clip extraction approach is applied on the testing set. To get the confidence score for a testing video, we apply a max pooling strategy over all the clips from that video.

References

1. Caba, F., Thabet, A., Niebles, J.C., Ghanem, B.: Camera movement and surrounding scene appearance as contextual features for action recognition. In: ACCV (2014)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
3. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV (2010)
4. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)