Combined Ordered and Improved Trajectories for Large Scale Human Action Recognition

O. V. Ramana Murthy[1]
[1]Vision & Sensing, HCC Lab,
ESTeM, University of Canberra
O.V.RamanaMurthy@ieee.org

Roland Goecke[1,2]
[2]IHCC, RSCS, CECS,
Australian National University
roland.goecke@ieee.org

Abstract

Recently, a video representation based on dense trajectories has been shown to outperform other human action recognition methods on several benchmark datasets. The trajectories capture the motion characteristics of different objects, for example human bodies, in spatial and temporal dimensions. In dense trajectories, points are sampled at uniform intervals in space and time and then tracked using a dense optical flow field over a fixed length time window of L frames (optimally 15) overlapping over the entire video. However, amongst these base trajectories, some continue for longer than duration L. These longer motion characteristics of objects may be more valuable than the information from the base trajectories or at least provide complementary information otherwise not captured. Thertefore, we propose a technique that searches for trajectories with longer duration and call these 'ordered trajectories'. We apply these ordered trajectories in conjunction with the recent 'improved trajectories' (improved dense trajectories) approach on the UCF101 dataset.

1. Ordered Trajectories

The overall layout of our proposed framework is shown in Fig. 1. Firstly, dense trajectories [4] are detected. The dense trajectories code available online[1] [4] was used in all our experiments. The 'ordered trajectories' have been recently proposed by [2]. They were proposed to select those trajectories from, the base dense trajectories, that have a longer duration. Dense trajectories are usually computed over every overlapping sequence of $L(=15)$ frames. However, some trajectories can continue for varying periods beyond the fixed length of L. The ordered trajectories technique generates all such trajectories by matching the dense trajectories of every two consecutive frames.

Local descriptors – Motion Bound Histograms (MBH), Histograms of Oriented Gradients (HOG), Histogram of Optical Flow (HOF) – of the selected matching trajectories are accumulated, while the *Trajectory Shape* descriptor is computed from the generated ordered trajectories as follows.

For a trajectory of given length L (number of frames) and containing a sequence of points $P_t = (x_t, y_t)$, the trajectory shape is described in terms of a sequence of displacement vectors $\Delta P_t = (P_{t+1} - p_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. The resulting vector is normalised by the sum of displacement vector magnitudes

$$T = \frac{\Delta P_t, ..., \Delta P_{t+L-1}}{\sum_{j=t}^{t+L-1} ||\Delta P_j||} \quad (1)$$

2. Improved Trajectories

Improved trajectories have been recently proposed and the code released by Wang *et al.* [5]. It is an improved version of the dense trajectories obtained by estimating the camera motion, which is estimated by matching feature points between frames using SURF descriptors and dense optical flow. The obtained matches are used to estimate a homography with RANSAC. Further, a human body detector is used to separate motion stemming from humans moving and from camera motion. The estimate is also used to cancel out possible camera motion from the optical flow. This technique has been shown to significantly improve motion-based descriptors such as MBH and HOF by removing apparent motion not related to the human body. In our experiments, we only use the camera motion compensated improved trajectories, without any human body detector. It is available online[2]

[1] http://lear.inrialpes.fr/people/wang/dense_trajectories

[2] http://lear.inrialpes.fr/people/wang/improved_trajectories

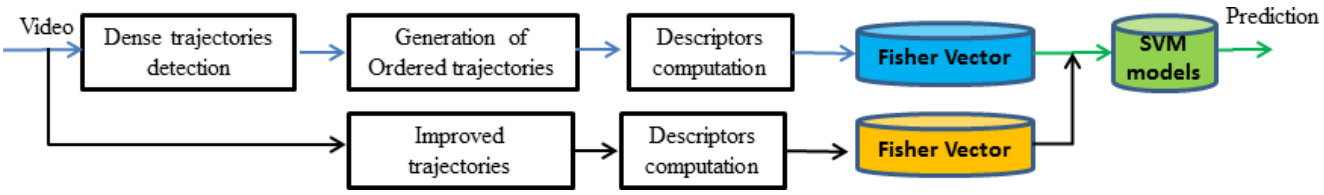

Figure 1: Dense trajectories are extracted from a video. Ordered trajectories are generated by matching dense trajectories of consecutive frames. Improved trajectories (without human body detector) are also extracted from the videos. Fisher vectors are constructed for each set of local feature descriptors of these trajectories. All Fisher vectors are concatenated and passed to a linear SVM for classification.

3. Fisher Vector Encoding

Most of the encoding techniques like the hard assignment capture only the information of frequency of the visual words (of the codebook) for given video. Fisher vectors capture the first order (deviations from the visual words) and second order (covariance deviation) statistics. Further, in an recent study on large scale image classification [1], fisher vectors have been found to perform best. So, we use Fisher vector encoding for constructing the features from the local descriptors. Firstly, 10,000 descriptors from each class of the training data are randomly selected. Accumulating these from the 101 classes of the UCF-101 dataset [3] roughly results in 10^6 descriptors. Next, Principal Component Analysis (PCA) is applied to reduce the descriptor dimensionality by a factor of two. Then, a Gaussian Mixture Model (GMM) is fitted to this data with the number of Gaussians $K = 256$. Each video is, then, represented by a $2 \times D \times K$-dimensional Fisher vector for each descriptor type, where D is the descriptor dimension after performing PCA. Finally, power normalisation and L_2 normalisation are applied to the Fisher vector as set forth in [1].

4. Results

The Fisher vectors obtained from the ordered trajectories and improved trajectories approaches are concatenated and passed to a linear SVM for classification of the human actions into the given 101 classes. The videos in 101 action categories are grouped into 25 groups, where each group can consist of 4-7 videos of an action. The videos from the same group may share some common features, such as similar background, similar viewpoint, etc. The Three train-test splits were provided for consistency. In each split, clips from 7 of the 25 groups are used as test samples, and the remaining for training. Recognition rates obtained for each split are shown in Table 1.

5. Conclusions

Dense trajectories have been found to perform the best on most of the existing human action recognition datasets.

Table 1: Split-wise classification

Splits	1	2	3	Average
Dense trajectories	81.72%	83.41%	82.91%	82.68%
Ordered trajectories	78.52%	81.08%	80.79%	80.13%
Improved trajectories	82.43%	84.22%	84.00%	83.55%
Improved + Dense	84.09%	**86.49%**	**85.54%**	85.38%
Improved + Ordered	**84.76%**	86.37%	85.18%	**85.44%**

However, they are usually computed over every set of $L(=15)$ consecutive frames only. This does not capture longer trajectories, which exist and provide important complementary information. From the base trajectories, we generate ordered trajectories that have a variable, longer duration than these base trajectories. Combining them with the improved trajectories approach, which is a recently proposed improved version of the dense trajectories approach with camera motion compensation, we hav applied them to the UCF101 dataset and achieved an improved performance over the baseline dense trajectories, the ordered trajectories and the improved dense trajectories.

References

[1] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on Computer vision (ECCV 2010)*, pages 143–156. Springer-Verlag, 2010. 4322

[2] O. V. Ramana Murthy and R. Goecke. Ordered Trajectories for Large Scale Human Action Recognition. In *IEEE Intenational Conference on Computer Vision Workshops (ICCV2013)*, Dec. 2013. 4321

[3] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild. In *CRCV-TR-12-01*. November 2012. 4322

[4] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 3169–3176, June 2011. 4321

[5] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *IEEE Intenational Conference on Computer Vision (ICCV2013)*, Dec. 2013. 4321