LPM for Action Recognition in Temporally Untrimmed Videos

Feng Shi, Robert Laganière and Emil Petriu
School of Electrical Engineering and Computer Scinece
University of Ottawa, Ottawa, On, Canada
{fshi098, laganier, petriu}@site.uottawa.ca

Abstract. In this notebook paper, we evaluate GBH and MBH descriptors for action recognition in temporally untrimmed videos. Our system is based on the recent improvement of local part model with gradient boundary descriptor [5]. We extract both local GBH and MBH descriptors and represent them with Fisher vector. We use LPM to include local structure information. We apply a slide window approach to extract short clips from temporally untrimmed video, and using a linear SVM to classify each short clip. We simply label the untrimmed video using the clip with the maximal classification score.

1 Introduction

Recent studies in human action recognition have achieved remarkable performance. Most of evaluations [5, 6, 9, 10] on action recognition in the literature use trimmed clips which are manually selected to bound the action of interest. The THUMOS14 workshop [3] aims to address such limitation by including temporally untrimmed videos. This notebook paper describes the VIVA-UnivOttawa submission on this challenge.

2 The methods

We use local part model (LPM) [8] to represent videos. We apply a slide window approach to extract short clips from temporally untrimmed video. We efficiently extract the local features with random sampling, and represent the LPM features with both gradient boundary descriptor (GBH) [5] and MBH [2] descriptors. We finally use improved fisher vector [4] to encode features, followed by a linear SVM for classification.

2.1 Review of LPM algorithm

Local part model was introduced by Shi *et al.* in [7]. Their original purpose was to address the orderless issue of the bag-of-features representation with overlapping local "parts". In addition to having the overlapping local part patches, the method also includes a coarse primitive level "root" patch which encodes local

global information. To improve the efficiency of LPM computation, two integral videos are computed, one for the root at half resolution, and another one for the parts at full resolution. The descriptor of a 3D patch can then be computed very efficiently through 8 additions multiplied by the total number of root and parts.

Later, Shi *et al.* [8] improved the efficiency by combining random sampling method with local part model. Random sampling does not require feature detection, which greatly improves processing speed. In this work, we treat the root and 8 parts as two separate channels. For each channel, a standard Fisher vector encoding is applied. The resulting Fisher vectors from root and parts are concatenated into one histogram for SVM classification. To better fit FV encoding, we reduce the descriptor dimensionality by using Principal Component Analysis (PCA). In our experiments, the dimensions of root descriptor and part descriptor are reduced into their 1/2 and 1/8, respectively.

2.2 Local feature descriptors

Local features and their descriptors have significant impact on the performance of the video recognition. In this paper, we use efficient GBH to encode local structure information, and state-of-the-art MBH descriptor to encode motion information.

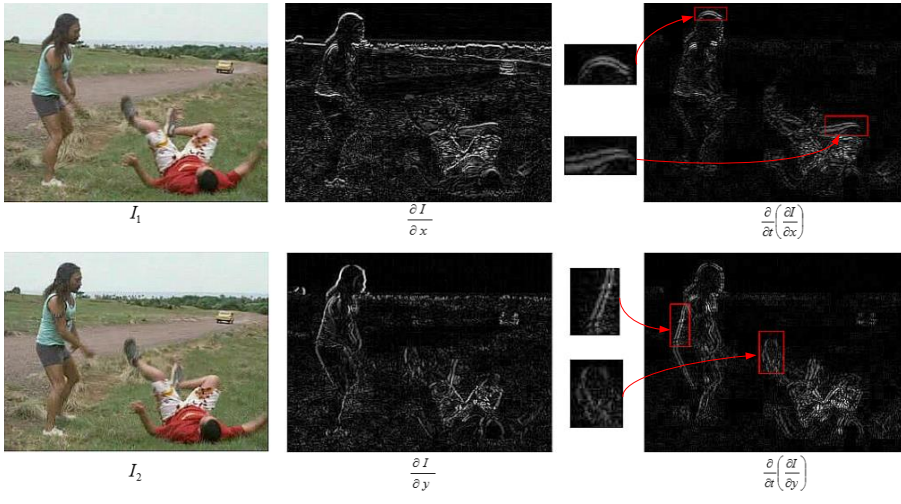

Fig. 1. Illustration of gradients and gradient boundaries for a "fall floor" action. Compared to image gradients, gradient boundaries have less background noise. More important, gradient boundaries encode motion information. The areas inside red bounding boxes show the double edges with various distances decided by the speed of the moving body parts.

Gradient boundary descriptor (GBH) is proposed in [5] for efficient action recognition. GBH uses a similar histograms of orientation based method voting with θ and r as in SIFT and HOG descriptors. However, instead of using image gradients, it uses time-derivatives of image gradients, which emphasize moving edge boundaries. By applying gradient subtraction, GBH avoids using the third gradient component as 3D SIFT and HOG3D descriptors which lead to high dimensionality and relatively expensive quantization cost. Instead, it adopts compact HOG-like descriptor with two gradient components.

Figure 1 shows the comparison of image gradients and gradient boundaries. It illustrates two important observations. First, the subtraction of two consecutive image gradients results in the removal of the backgrounds of the video sequences. The two gradient images in the centre show a lot of background noise, while the gradient boundary images on the right show clear human shapes with far less background noise. More important, gradient boundaries encode the moving human shapes. As demonstrated with red bounding boxes in the figure, the double edges with various distances are proportional to the moving speed of the human body parts. For example, the distance between double leg edges is larger than the double head edges, which represents that the leg moves faster than the head of the other person in the upper right image.

2.3 Feature encoding

We strictly follow the evaluation methods as [5], and use its codes[1] to compute both GBH and MBH descriptors. We randomly extracted LPM features from the video, and compute GBH and MBH descriptors with exactly the same parameters as [5]. For MBH descriptor, we compute optical flow with duality-based TV_L1 approach [11], which estimates more accurate and discontinuity-preserving flow field. We use improved fisher vector [4] by applying the signed square-rooting and followed with L_2 normalization.

2.4 Parameters

We use random sampling to extract 10000 root patches from the root video. For each root patch, we sample 8 ($2 \times 2 \times 2$) overlapping part patches from the part video. The histograms of 1 root patch and 8 part patches are treated as two separate channels. We use minimal patch size of $20 \times 20 \times 14$. The consecutive scales are computed by multiplying the patch by a factor of $\sqrt{2}$. With total of 8 spatial scales and 2 temporal scales, we sample a video 16 times. Each patch is subdivided into a grid of $2 \times 2 \times 2$ cells, with no sub-block division. 8 bins are used for quantization, which leads to a feature dimension of 64. Thus, a LPM feature (MBHx, MBHy or GBH) has a root channel of dimension 64 and a part channel of dimension 512. We set the number of Gaussians to K = 128 and randomly sample a subset of 150,000 features from the training set to estimate both GMM and PCA projective matrix. We first use PCA to reduce root feature from 64

[1] https://github.com/fshi

to 32 and part feature from 512 to 64. Then we encoding them into FVs as two channels.

2.5 Sliding window approach

For training, we only use all 13,320 clips from the temporally trimmed UCF101 [9] dataset. We first compute FVs for both GBH and MBH, and then concatenate them to feed in a linear SVM implemented by LIBSVM [1] with $C = 32.5$. The training is performed with probability models, which provide the probability estimates for testing. For multiple classes, we use one-verse-one approach.

For testing on a temporally untrimmed video, we apply a sliding window approach to extract short clips, and using the learnt SVM models to classify each short clip. We set the sliding window length as 160 frames, and slide the temporal window at a step of 50 frames. For each sliding window, we compute FVs for both GBH and MBH and use the learnt SVM probability models to predict its class probability. After performing prediction on all sliding windows, we simply label the untrimmed video using the window with the maximal classification score. We also report the average score of three sliding windows, which show larger classification scores than other windows.

For fast processing, we down-sample both the UCF101 and the temporally untrimmed testing videos to half size, then use lower resolution videos (160x120 for UCF101 and 160x90 for untrimmed videos) to compute FVs for both GBH and MBH.

References

1. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3), 27 (2011)
2. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV, pp. 428–441. Springer (2006)
3. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/ (2014)
4. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV (2010)
5. Shi, F., Laganiere, R., Petriu, E.: Gradient boundary histograms for action recognition
6. Shi, F., Laganiere, R., Petriu, E., Zhen, H.: Lpm for fast action recognition with large number of classes. In: THUMOS: ICCV Workshop on Action Recognition with a Large Number of Classes (2013)
7. Shi, F., Petriu, E.M., Cordeiro, A.: Human action recognition from local part model. In: Proc. IEEE Int Haptic Audio Visual Environments and Games (HAVE) Workshop. pp. 35–38 (2011)
8. Shi, F., Petriu, E., Laganiere, R.: Sampling strategies for real-time action recognition. In: Proc. IEEE Conf. Computer Vision Pattern Recognition. IEEE (2013)
9. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. Tech. Rep. CRCV-TR-12-01, CRCV, University of Central Florida (2012)
10. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: International Conference on Computer Vision (2013)
11. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: Ann. Symp. German Association Pattern Recognition. pp. 214–223 (2007)