Human Action Recognition Using Scene Context

Qiuwei Shou
Indiana University
qiuwshou@umail.iu.edu

1. INTRODUCTION

Human action recognition has been studied widely in computer vision.Many researches have been done on static camera and surveillance to recognize action in past decades. Recently more and more approaches have been studied to capture actions in dynamic videos. It becomes a more challenging and interesting task to detect or recognize action in highly unconstrained dataset. Many state-of-the-art methods developed by deep learning approaches have shown good performance on web datasets like UCF50 and HMDB51. Image based model also have been improved by taking account the scene information.

There are various approaches to combine the scene information like using CRF and super pixels. In this project, we will try to develop a image based model by applying CSIFT[7] to capture scene context. The project is inspired by [6]. We use a similar architecture and different setup which will be describe in section 4. The reference implementations are color sift [7] and opencv[1]. In section 2, we will briefly describe the problem and scene context descriptor, in section 3 we will describe our implementation of applying CSIFT and motion descriptors, in section 4 we will evaluate our method using UCF50 dataset[2] and we conclude this project in section 5.

2. SCENE DESCRIPTORS

Scene context is useful to put constrain on the actions. For example, skiing and skate boarding have similar motion patterns, but these actions take place in different context and scene. Skiing happens on snow, which is very different from where skate boarding is done[6]. Many actions are associated with particular scene and context. Many researches have discovered the dependency of particular actions upon certain scene using probabilistic model [5]. It have been shown that combing scene context can increase the accuracy of classification. We define the scene context as stationary pixels in a image. We compute the dense optical flow (u, v) at each pixel by using Gunner Farneback's algorithm[4]. We use CSIFT[7] to distinguish the scene context. For example, skiing usually happens in the snow environment which will generate distinct CSIFT descriptor.

3. PROPOSED APPROACH

We define the scene context as stationary pixels in a image. For each video clip, we sample 4 key frames evenly through the whole video. Using more key frames will increase the accuracy with increasing the run time (see figure 1). We compute the dense optical flow (u, v) at each pixel between two consecutive frame, using Gunner Farneback's algorithm[4]. Apply in a threshold to the displacement, we decide if a pixel is moving or stationary. We compute the CSIFT for the whole image then separate pixels into stationary group SP and moving group MP by pixel index. We extract CSIFT from both group and use BOW paradigm to create two histogram representations of 200 codebook (see figure 2). We also concatenate all the dense optical flows for consecutive key frames and create a histogram representation M with the same number of codebook for each video clip. We first train the scene context descriptor and motion descriptor separately by SVM classifier[3]. Then we use early fusion to combine the two feature descriptors together. See figure 2 for the work flow.

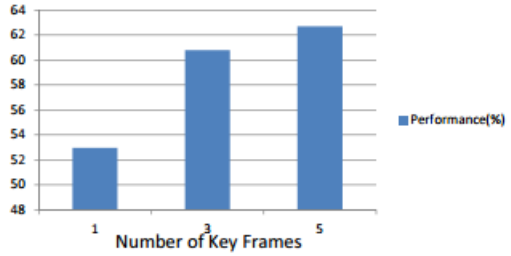

Figure 1: Performance vs Number of key frames [6]

4. EVALUATION

In this section we will conducting evaluation of our method using UCF50 dataset [2]. UCF50 is a collection of web videos with 50 categories, and we will use our method to classify the category of videos. Under each category there are 25 groups of video. The video clips in the same group are taken from the same video. We split data into training set and test set half by half evenly through the group to guarantee we have the training data and test data from the same video. by key frames of 4 we totally sampled 26724 images of 240*320 dimensions. For each one we extact 1957 CSIFT descriptors of 384 dimensions vector by dense sampling. We use 2 for threshold to distinguish the moving pixel and stationary pixel. Averagely using the motion descriptor we achieve about 48% accuracy for only using motion descriptor and 45% by only using scene context descriptor. Combing two descriptors achieve overall performance of 54% . In the [6]

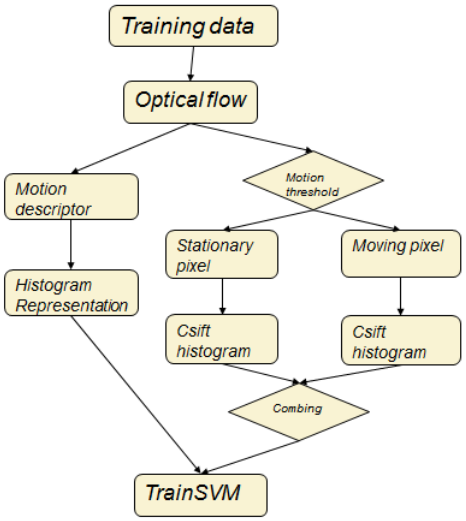

Figure 2: Proposed Approach

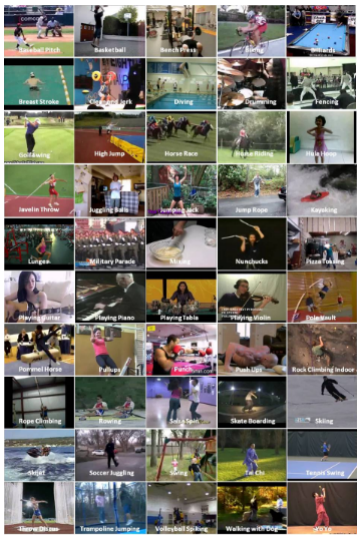

Figure 3: A snapshot of UCF50 dataset

the performance for using motion descriptor and scene descriptor are 53% and 49%, and late fusion of both gives overall performance of 68% with 1000 dimension codebook and 3 key frames. So we didn't achieve large improvements by early fusion. Because the original experiment details aren't fully explained in the [6] and there are many parameters need to be tuned. There more details could be explored to increase the performance. According to the results in the figure 4, it shows that the experiment setup can affected the performance. Also the threshold value might be critical. If it is too small we might lose context information for the object and action, or if it is too larger we can lost scene context information. The window stride and pyramid scale can affect the performance of dense optical flow. Increase key frames and dimension of codebook definitely increase the performance. Also the assumption is not flawless because we assume that the scene is homogeneous. For the actions that could happen in different scenes it might lose the accuracy. Moreover, the length of video clip is various. It causes the ambiguity to produce the same motion patterns for the video clip if the frames are sampled at differ fps rate. Splitting the data half by half might give us less training data so the classifier is not learned well. Also we use 10 iteration and 0.1 epsilon for generating codebook, which might not be converged.

Performance	Experimental Setup	Paper
76.90%	Leave One Group Out Cross-validation (25 cross-validations)	Reddy and Shah. (MVAP), 2012
57.90%	5-fold group-wise cross-validation	Sadanand and Corso. (CVPR), 2012
76.40%*	Video Wise Cross-validation (*Since videos belonging to a group are obtained from a single long video, similar videos can end up in both training and testing in "video-wise cross-validation" leading to high performance)	Sadanand and Corso. (CVPR), 2012
81.03%*	2/3 training and 1/3 testing for each class (*From the details given in the paper, we are not sure if videos belonging to the same group are kept seperate in training and testing sets and the paper does not give details on number of cross-validations)	Todorovic. (ECCV), 2012
73.70%	Leave One Group Out Cross-validation (25 cross-validations)	Solmaz, et al. (MVAP), 2012
72.60%	Leave One Group Out Cross-validation (25 cross-validations)	Kliper-Gross, et al. (ECCV), 2012

Figure 4: Other results on the UCF50 dataset

5. CONCLUSION

We didn't achieve the good performance with higher key frames than the [6]. But we have proved that combing scene context definitely help to increase the accuracy of the classification by 6%. Using image based model doesn't require large amount data set and can work for unconstrained data well. But it can hit the boundary easily. As the deep learning is the big trend in the computer vision, for the future enhancement, we want to apply CNN for scene classification with more appropriate descriptors.

6. REFERENCES

[1] Opencv. http://opencv.org/.

[2] Ucf50. http://crcv.ucf.edu/data/UCF50.php.

[3] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[4] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pages 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.

[5] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[6] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Mach. Vision Appl.*, 24(5):971–981, July 2013.

[7] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.