ACCELERATED LOCAL FEATURE EXTRACTION IN A REUSE SCHEME FOR EFFICIENT ACTION RECOGNITION

Jia-Lin Chen, Zhi-Yi Lin, Yi-Chen Wan, Liang-Gee Chen

National Taiwan University
Graduate Institute of Electrical Engineering
DSP/IC Design Lab

ABSTRACT

In this paper, we propose an accelerated local feature extraction in a reuse scheme for action recognition. Most local features of the previous frame could be reused due to the high correlation between successive frames. Feature extraction is only needed to be applied partially in the current frame. The full-frame features of the current frame are combined by features extracted at different times. Experimental results show that the proposed reuse scheme can achieve comparable performance to full-frame feature extraction while computational effort is reduced significantly.

Index Terms— Feature reuse, local feature extraction, partial feature extraction, action recognition, region of interest

1. INTRODUCTION

Action recognition is an important research topic in the computer vision community and has received more and more attention in recent years. It is due to the dramatically increasing amount of videos and the far-reaching applications including intelligent surveillance, smart home care monitoring systems, content based video retrieval, automatic sports analysis, and humancomputer interaction. In order to provide real-time recognition results for immediate reaction, a fast feature extraction algorithm is strongly demanded.

In video sequences, action can be represented by global features or local features. Global features are extracted using the entire frame information, while the local features are extracted from the local portion of the whole frame. Recently, several local features for action recognition have been proposed, including histograms of flow orientations (HOF) [1], 3D SIFT[2], histograms of 3D gradients (HOG3D)[3], motion boundary histograms (MBH) [4], shapes of point trajectories [5], local trinary patterns [6] and others. In literature, the work demonstrates that MBH, HOF and HOG descriptors sampled along dense point trajectories outperform other methods on a number of challenging datasets [4].

However, the local features are usually densely extracted from the whole frame which is high computation and time-consuming. In real scenarios, in order to provide immediate reactions, the action recognition has to be completed in small amount of time under limited computation resources and power. It is noticed that there is only partial information useful for further recognition in most of the cases. The concept of Region of Interest (ROI) is often adopted to determine a region needed to apply feature extraction and avoid full-frame feature extraction. But the correlation between successive frames is rarely considered.

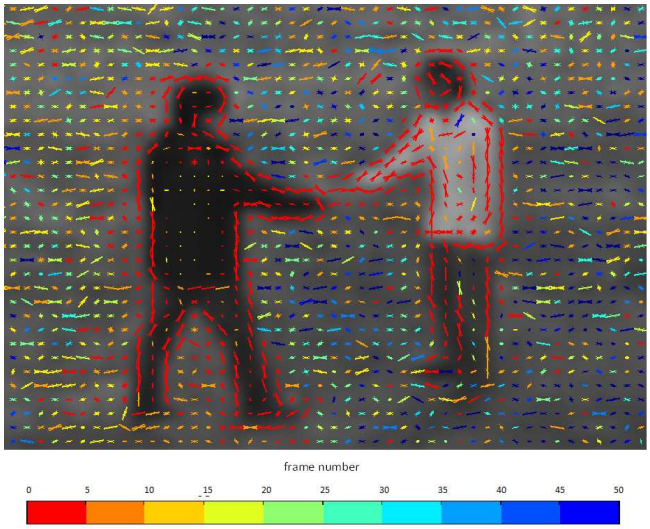

Fig. 1. An example of feature extraction results in the proposed reuse scheme. The color indicates the ”age” of the local feature. And the full-frame features are combined by features extracted at different times.

In this paper, we propose an accelerated local feature extraction algorithm in a reuse scheme. The method is not only applying feature extraction partially in ROI, but also reusing the features already obtained formerly. Since the successive frames are similar to a certain extent, a majority of the features in the previous frame could be reused when describing the current frame. Feature extraction is only needed to perform in partial regions. Combining with features reused from

previous frame, the full-frame features in current frames can be obtained while saving computational effort.

The main contributions of this work can be summarized as follows. (1) We propose an accelerated local feature extraction in a reuse scheme, which makes best use of the high correlation between successive frames; (2) we demonstrate action recognition using feature proposed; (3) we discuss the effect of feature reuse and show that the full frame features consisted of features extracted in different time are sufficiently useful.

The remainder of this paper is organized as follows. In the following section, we introduce the accelerated feature extraction in a reuse scheme. The action recognition with proposed feature is described in section 3. Section 4 shows our experiment results and analysis. Section 5 concludes this paper.

2. ACCELERATED FEATURE EXTRACTION IN A REUSE SCHEME

Given two successive frames of a video and the local features of the previous frame, our goal is to generate full-frame features of the current frame by reusing previous features and partial feature extraction in ROI.

There are two computation units in the local feature extraction discussed in this paper. One is the cell and the other is the block. The size of the cell is 8 pixels by 8 pixels and one block consists of four cells. Our approach mainly involves three steps: ROI determination (Subsection 2.1), partial feature extraction in ROI (Subsection 2.2), and feature combination with the previous features (Subsection 2.3). The details of each step are described below.

2.1. Region of Interest Determination

The ROI determination is operated on a cell to coordinate the following steps. Each color frame I is downsampled by taking average cell-wise as shown below

$$I'(u,v,ch) = \frac{1}{cellsize} \times \sum_{(x,y)\in c'} I(x,y,ch),\ ch \in \{R,G,B\} \tag{1}$$

where c' is a standard 8 pixels by 8 pixels cell and *cellsize* is 64 in our setting. The difference map D is the sum of absolute differences on three color channels and used to measure the change between the downsampled previous frame I'_{pre} and downsampled current frame I'_{cur}.

$$D(u,v) = \sum_{ch} \left| I'_{cur}(u,v,ch) - I'_{pre}(u,v,ch) \right| \tag{2}$$

The ROI map is defined as

$$ROI(u,v) = \begin{cases} 1 & \text{if } D(u,v) > \tau \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The threshold τ is the r-th percentile of the total elements in difference map, where r is a ratio of ROI area to the whole frame area. A smaller r corresponds to a stricter detector, which means that less cells will be selected for feature extraction.

2.2. Local Feature Extraction

The local features considered in this work are HOG, HOF and MBH. All their computation unit is the cell and the cell feature extraction can be divided into the two steps, vector computation and orientation binning.

2.2.1. Vector Computation

In the first step, two-dimensional vectors are computed for each local feature respectively. For HOG in color images, the gradient is computed separately for each color channel. The one with largest magnitude is used as following

$$\begin{gathered} \mathbf{v_{HOG}} = [\ G_x(I(x,y,\tilde{ch}))\ G_y(I(x,y,\tilde{ch}))\ ], \\ \tilde{ch} = \underset{ch\in\{R,G,B\}}{\arg\max} \sqrt{G_x(I(x,y,ch))^2 + G_y(I(x,y,ch))^2} \end{gathered} \tag{4}$$

and $G_x(.)$ and $G_y(.)$ are to apply the 1-D centered, point discrete derivative mask to input data P in horizontal and vertical directions separately.

$$\begin{aligned} G_x(P(x,y)) &= P(x+1,y) - P(x-1,y) \\ G_y(P(x,y)) &= P(x,y+1) - P(x,y-1) \end{aligned} \tag{5}$$

For HOF, the vector used is the optical flow, while for MBH is the gradient of optical flow magnitude.

$$\mathbf{v_{HOF}} = [\ OptFlow(x,y,1)\ OptFlow(x,y,2)\ ] \tag{6}$$

$$\mathbf{v_{MBH}} = [\ G_x(\|OptFlow(x,y)\|)\ G_y(\|OptFlow(x,y)\|)\ ] \tag{7}$$

2.2.2. Orientation Binning

Then next step is to create the cell histograms. Each pixel within the cell casts a weighted vote for an orientation-based histogram channel. There are 9 channels in HOG and MBH histogram, which spread over 0 to 180 degrees in an unsigned manner. Since the orientation of the optical flow is an important cue, there are 18 channels in HOF histogram channels spreading over 0 to 360 degrees in a signed manner. The

contribution of each vector **v** to the histogram is given by its magnitude and equally split to the two closest bins.

$$m_{\alpha_1} = 0.5 \times \|\mathbf{v}\| \\ m_{\alpha_2} = 0.5 \times \|\mathbf{v}\| \tag{8}$$

Instead of computing the vector orientation by arctangent function, the closest bin is found by the inner product of **v** and the unit vector $\boldsymbol{\theta}_\alpha$ pointing at the direction α-th bin centered at.

$$\alpha_1 = \underset{\alpha \in A}{\arg\max}\ \boldsymbol{\theta}_\alpha \cdot \mathbf{v}^T \\ \alpha_2 = \underset{\alpha \in A \setminus \{\alpha_1\}}{\arg\max}\ \boldsymbol{\theta}_\alpha \cdot \mathbf{v}^T \\ \boldsymbol{\theta}_\alpha = [\cos\theta_\alpha \ \ \sin\theta_\alpha] \tag{9}$$

2.3. Feature Combination

The cell histogram feature in current frame h_{cur} combines the features from the previous frame h_{pre} and the new extracted features h_{new} in ROI as shown below

$$h_{cur}(u,v) = ROI(u,v) \times h_{new}(u,v) \\ + (1 - ROI(u,v)) \times h_{pre}(u,v) \tag{10}$$

After grouping cells into blocks, the cell histogram features are normalized based on all histograms in the block. The blocks are 2 cells by 2 cells and have 50% overlap, which could provide tolerance for dealing with cell features extracted at different time. Finally, the full-frame features are represented as the concatenation of HOG, HOF, and MBH.

3. ACTION RECOGNITION WITH COMBINED FEATURES

To evaluate the effect of feature reuse, we demonstrate action recognition using features extracted in the proposed reuse scheme. In order to obtain the video representation, we follow the work in [7] and try to discover the discriminative parts in a weakly-supervised manner.

Firstly, we sample patches with sufficient texture and motion information from training videos. The extracted patches are described with HOG, HOF and MBH. Then a standard k means is performed in the whitened feature space [8] to obtain a large collection of clusters of patches. A universal negative model from all patches is learned once to accelerate the exemplar-SVM method. Then an cross-validation procedure is adopted iteratively to prune noisy examples and refine effective detectors [9]. The distinctive action part detectors are selected by coverageentropy curves and similarity measurement to measure the capability of a part detector and the similarity between two detectors [7].

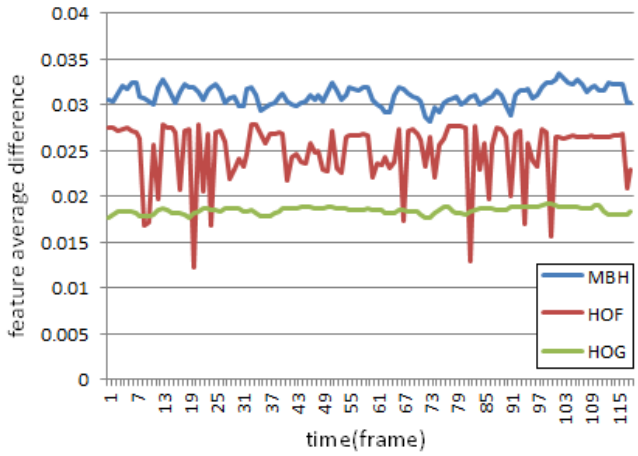

Fig. 2. Feature average difference versus time.The difference is oscillating but not increasing dramatically as times go by.

All part detectors discovered are used to construct mid-level representation for videos. We first run each detector in a sliding 3D-window fashion on videos. The highest detection score of each detector is concatenated as the final feature vector of a video. A standard SVM classifier[10] is used on the obtained feature vector and the one-versus-the-rest scheme is adopted for multi-class action classification.

Since the training process often applies off-line and has smaller real-time demand, the local features are all extracted without reusing in the training stage. While in testing stage, the recognition usually should be completed in small amount of time and limited resources, the frames are represented by the feature extracted in the proposed reuse scheme.

4. EXPERIMENTS

4.1. Analysis of Proposed Features

The proposed accelerated local feature extraction is in a reuse scheme. Feature extraction is only needed to perform partially in ROI. Combining with features from previous frame, the full-frame features in the current frame can be obtained. After applying the proposed feature extraction frame by frame, the features of a frame would be consisted of features extracted in different time, and the difference between the accelerated feature extraction and full-frame feature extraction may be increasing with time.

An example of HOG feature extraction result in the proposed reuse scheme is shown in fig. 1. Standard HOG features with a cell size of eight pixels are drawn in different color. The HOG of each cell accumulated from a sequence of 47 frames using our proposed method. The color indicates the ”age” of the cell HOG feature in denomination of frame number. The feature drawn in blue means it is extracted tens of frames ago. And the features drawn in red are extracted

from the last few frames. It is noted that cells with larger appearance changes, like the boundary of an action, are always being updated. In contrast, cells with less appearance changes like background would not be updated frequently. And thus the proposed feature extraction could provide sufficient information in an efficient way.

Fig. 2 shows the feature average difference versus time. The features are normalized with blocks already. The difference is computed by average of absolute differences between the features extracted in a reuse scheme and the full-frame extracted feature in the current frame. The blue, red and green curves represent the difference of MBH, HOF and HOG respectively. It is noticed that although there are some error compared to the full-frame extracted feature, the difference is oscillating but not increasing dramatically as times go by.

4.2. Results of Action Recognition

We validate the proposed system on the UT-Interaction benchmark dataset [11] using the segmented versions for learning and detection setting. The segmented version of the dataset contains one human-human interaction in each video. There are a total of 120 videos including 6 classes of interactions: handshake, point, hug, push, kick and punch.

Based on the 10-fold leave-one-out cross-validation, our method is evaluated and shown in table 1. According to the average action recognition accuracy, the proposed method using 10% as ROI ratio, which means only perform feature extraction in 10% of the whole frame area, can achieve 86.7% recognition accuracy and reduce 90% computational effort. And when perform feature extraction in 30% of the whole frame area can reach 88.3% recognition accuracy, which is the same as using full-frame extracted features. The accuracy of our system is lower than the state-of-the-art performance, but it shows that the proposed combined feature can provide sufficient information.

Table 1. Average Recognition Accuracy versus ROI Ratio

ROI Ratio (%)	10	30	50	70	90
Accuracy (%)	86.7	88.3	88.3	88.3	88.3

5. CONCLUSION

In this paper, we propose an accelerated local feature extraction in a reuse scheme for action recognition. Most local features of the previous frames could be reused due to the high correlation between successive frames. Feature extraction is only needed to be applied partially in the current frame, which greatly reduces the computation load. The full-frame features of the current frame are combined by features extracted at different times. Experiments show that the combined features can provide sufficient information since the regions with larger appearance changes are always being updated. The action recognition with combined features can achieve comparable performance to which using full-frame extracted features.

6. REFERENCES

[1] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 428–441.

[2] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 357–360.

[3] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[4] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[5] P. Sand and S. Teller, "Particle video: Long-range motion estimation using point trajectories," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 72–91, 2008.

[6] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 492–497.

[7] F. Chen, N. Sang, X. Kuang, H. Gan, and C. Gao, "Action recognition through discovering distinctive action parts," *JOSA A*, vol. 32, no. 2, pp. 173–185, 2015.

[8] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative decorrelation for clustering and classification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 459–472.

[9] S. Singh, A. Gupta, and A. Efros, "Unsupervised discovery of mid-level discriminative patches," *Computer Vision–ECCV 2012*, pp. 73–86, 2012.

[10] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[11] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1036–1043.