A Wavelet Based Local Descriptor for Human Action Recognition

Ling Shao
ling.shao@sheffield.ac.uk
Ruoyun Gao
ruoyun.gao@gmail.com

Department of Electronic and Electrical Engineering,
The University of Sheffield
Department of Computer Science,
Leiden University

In interest point based human action recognition, local descriptors are used to represent information in the neighbourhood around each extracted space-time interest point. The performance of the action recognition systems highly depends on the invariance and distinctiveness of the local spatio-temporal descriptor adopted. In this paper, we propose a new descriptor based on the Wavelet Transform taking advantage of its capability in compacting and discriminating data. We evaluate this descriptor on the extensively studied KTH action dataset, using the Bag-of-Features framework. Results show the Wavelet Transform based descriptor achieves the recognition rate of 93.89%, which is better than most of the state-of-the-art methods.

Recognizing human actions is a challenging problem, which needs to deal with various variations of people's posture and clothes, dynamic background, and partial occlusion, etc. Intuitively, a straightforward way is comparing an unknown video with the training samples by computing correlation between the whole videos [1]. This approach makes good use of geometrical consistency, but it is not feasible in dealing with camera motions, zooming and non-stationary background, etc.

To conquer these deficiencies, a lot of researchers focus on part-based approaches for which only the 'interesting' parts of the video are analyzed other than the whole video. These mentioned 'parts' can be trajectories or flow vectors of corners, profiles generated from silhouettes and spatial temporal interest points. We adopt the recognition scheme that utilizes spatial temporal interest points, which are robust to viewpoint variation, moving background, zooming and object deformation [2]. In this framework, we extract feature descriptors from all the training sequences and build a codebook by clustering similar features. The cluster centers are the members of this codebook named as video words. Each feature descriptor is assigned to a certain video word. In this way, it simplifies the usage of all the extracted features while still maintains their richness in information [3]. This way of representation is known as 'Bag of Visual Words', which ignores the geometric arrangement between visual features and an action video is represented as a histogram of the number of occurrences of particular video words [4]. Then classification methods are exploited to build models for each action class and give the predicted class label for each test sequence. Numerous discriminative classifiers have been well exploited, such as Support Vector Machine (SVM), K Nearest Neighbour (KNN), and generative models such as Probabilistic Latent Semantic Analysis (pLSA) [4] and Linear Discriminant Analysis (LDA).

The aim of this paper is to propose a new feature description method to make features more reliable in representing sequences and further improve the recognition rate. Based on the wide usage and good performance of time frequency transformation techniques in the image processing area, we introduce the Wavelet Transform, which is compact and informative, into the human action recognition application.

After spatio-temporal interest points (cuboids) are detected, description methods are applied on the detected cuboids to represent information inside the cuboids. In the following, we describe the proposed Wavelet transform based descriptor.

$$X \xrightarrow{T} \tilde{X} \xrightarrow{process} \tilde{Y} \xrightarrow{T^{-1}} Y$$

Fig. 1: General signal transformation process.

The general process of signal transformation is given in Fig. 1. X is any signal, and T is some transform technique and *'process'* in the middle is usually non-linear. T^{-1} is the reverse transformation. Simply speaking, transformation techniques are to convert a signal to different frequency components. There are many well explored transformation techniques in the signal processing field, such as the Fourier Transform and the Wavelet Transform.

Fig. 2 illustrates the structure of one single cuboid. For each cuboid, we select three orthogonal planes intersecting in the middle of the cuboid and apply 2D Wavelet transform on these planes. The transformed coefficients are the discriminative and reliable information used as the feature descriptor for a video cuboid.

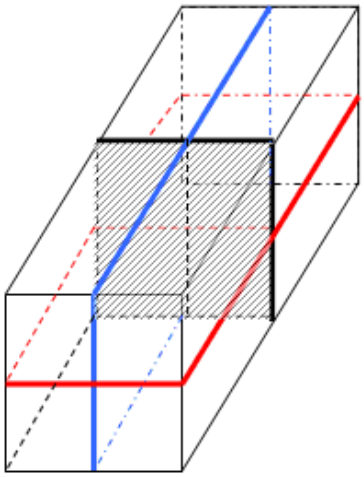

Fig. 2: Illustration of a cuboid and its internal three orthogonal planes.

Unlike the Fourier Transform, which utilizes just the sine and cosine functions, Wavelet Transforms have an infinite set of possible basis functions. Thus Wavelet analysis provides immediate access to information that can be obscured by other time-frequency methods such as Fourier analysis. For instance, in order to isolate signal discontinuities, one would like to have some very short basis functions. Meanwhile, in order to obtain detailed frequency analysis, one would like to have some very long basis functions. One way to achieve this is to have short high-frequency basis functions and long low-frequency ones. Wavelet Transforms gives exactly what we expect. Therefore, the adoption of multiple basis functions of wavelets gives perfect representation of the local information.

We compare our method to the state-of-the-art on the KTH dataset. The experimental results validate its reliability and efficiency in human action recognition, and also indicate the potential of transformation techniques in spatio-temporal video event analysis. The Wavelet Transform based method, although simple and short in feature length, outperforms the state-of-the-art action recognition algorithms significantly.

[1] S. Wong, T. Kim and R. Cipolla. Learning motion categories using both semantic and structural information. *In Procedings of IEEE Conference on Computer Vision and Pattern Recognition*. pp.1-6, 2007.

[2] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. *In Procedings of IEEE International Conference of Computer Vision*. Vol. 2, pp.1470-1477, 2003.

[3] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie. Behavior recognition via sparse spatio-temporal features. *In the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. pp. 65-72, 2005.

[4] J. C. Niebles, H. Wang and Fei-fei Li. Unsupervised learning of human action c-ategories using spatial-temporal words. *International Journal of Computer Vision*. Vol. 78, pp. 299-318, 2008.