A Wavelet Based Local Descriptor for Human Action Recognition

Ling Shao[1]
ling.shao@sheffield.ac.uk
Ruoyun Gao[2]
ruoyun.gao@gmail.com

[1] Department of Electronic and Electrical Engineering, The University of Sheffield, UK
[2] Department of Computer Science, Leiden University, The Netherlands

Abstract

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.

1 Introduction

Nowadays, more and more people record their daily activities using digital cameras, and this brings the enrichment of video sources on the internet, and also causes the problems of how to categorize existing video sources and how to classify new input videos according to their action classes. Apparently, categorizing these videos for future processing is a time-consuming task if it is done manually, and recognizing certain actions from scenes of interest in real movies is impossible to accomplish by manual effort. That is why these problems attract a lot of researchers' attention. They attempted to build a pattern recognition system, which trains on the feature descriptors extracted from the training videos and enables the computer to identify the actions of new videos automatically. The objective of this paper is to bring transform based descriptors into the human action recognition area and to validate their efficiency.

The development of computer vision has encouraged the occurrence of different novel recognition methods in both 2D images and 3D video sequences. Although it is still challenging to recognize a specific object from a dataset of images due to viewpoint changes, illumination, partial occlusions, and intra-class difference and so forth, many successful methods have been proposed [1, 2, 3]. But for the video recognition problem, the

BMVC 2010 doi:10.5244/C.24.72

current methods still need improvement, especially for recognizing human actions in real movies which have more individual 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 [4]. 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 [1]. 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 [5]. 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 [6]. 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) [6] 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.

The rest of the paper is organised as follows. Section 2 reviews the research background and the state-of-the-art methods in human action recognition. We illustrate the framework of human action recognition and introduce the new descriptor in Section 3. In Section 4, we apply the Wavelet descriptor on action recognition and evaluate its performance. Finally, we conclude in Section 5.

2 Related Work

Piles of work have been done in human action recognition and localization to save manual work and accelerate processing efficiency. Although there is no perfect algorithm to deal with all the problems occurring in action recognition, such as viewpoint changes, complex background activities and partial occlusions, yet the current work has shown quite promising results under different scenarios. One branch of work is to utilize the global information of the subject subtracted from video data, for example the correlating optical flow measurements from low resolution videos proposed by *Efros et al.* [7], which segment and stabilize each human figure and annotate each action in the resulted spatial temporal volume. Another part of work attempts to track the body parts and use the motion trajectories to discriminate different actions [8, 9]. In their implementations, certain feature points are located in a frame-by-frame manner, and the tracks of these points show many discriminative properties, such as position, velocities and appearance. However, the methods mentioned above are sensitive to partial occlusion and use much redundant information that is computationally expensive. The drawbacks of these methods accelerate the prosperity of part-based approaches, especially the space-time interest point detectors proposed by *Laptev et al.* [10] and *Dollár et al.* [5]. *Laptev et al.* [10] propose a space-time interest point detector based on the idea of the *Harris* and *Förstner* interest point operators. Their approach detects local structures in space-time where the image values have significant local variations in both space and time. *Dollár et al.* [5] use a set of separable linear filters detecting interest points with strong motion. This method is designed to respond to complex motion of local regions and the space-time corners. It detects more number of interest points than *Laptev*'s approach, which makes it more reliable in processing videos with limited frames. *Ke et al.* [11] apply spatial temporal volumetric features that efficiently scan video sequences in space and time and detect interest points over the motion vectors.

Transform domain techniques have been widely used in the image processing field, such as image compression, enhancement and segmentation etc. There are three well known transforms in this area, which are Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). Fourier Transform is to analyze a signal in the time domain for its frequency content and translate it into a function in the frequency domain. DFT estimates the Fourier transform of a function from a finite number of its sampled points [12]. DCT is similar to DFT, but only uses real data with even symmetry. DWT decomposes a discrete signal into a set of discrete basis functions (wavelets) instead of sine and cosine waves used in Fourier Transform, and it captures both the frequency information and the space information.

Recently, DCT and DWT have been widely exploited by modern image and video coding standards, such as JPEG, MPEG etc. [13, 14]. *Smith et al.* [15] propose a texture classification method based on the variance and the mean absolute values of the DCT coefficients calculated over the entire image. *Climer et al.* [16] propose a quadtree-structure-based method using the DCT coefficients on the nodes of the quadtree as image features. On the other hand, Wavelets have been used in detecting salient points

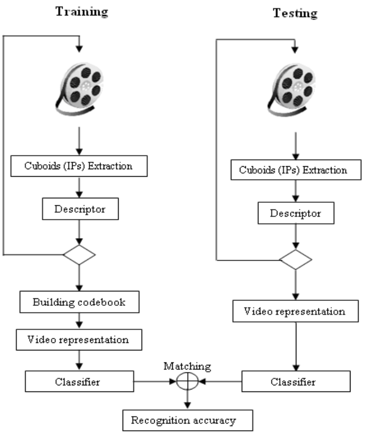


Fig. 1: Flowchart of the algorithm. IP is short for Interest Points.

representing local properties of images in content-based image retrieval [17]. *Schneiderman et al*. [18] describe a statistical method for 3D object detection by histogramming the subset of Wavelet coefficients and their position on the object. *Gonzalez-Audicana et al.* [19] use the multiresolution Wavelet decomposition to execute the spatial detail information extraction for image fusion.

The primary advantage of these transformation methods is the removal of redundancy between neighbouring pixels, and it leads to uncorrelated transform coefficients which can be encoded independently. Among all the transform techniques, Wavelet Transform has the best performance by localizing in frequency and also in space. Therefore Wavelet Transforms are desirable to deal with signals which are localized in time or space (speech or imagery) [20, 21].

3. Human Action Recognition Framework

The whole recognition process can be divided into two phases: the training phase and the testing phase. During the training phase, as the flow chart shown in Fig. 1, the interest points as well as the cuboids surrounding them are extracted by some interest point detector from the training sequences, and then descriptors of each sequence are generated by the structural distribution of interest points or the appearance information embedded in each cuboid. Descriptors from all training sequences are gathered together for further clustering by K-means which uses Euclidean distance as the clustering metric. The cluster centres are represented as the video words and they constitute the codebook. Each feature descriptor is assigned to a unique video word based on the distance between the descriptor and cluster centres. And the codebook membership of each feature descriptor is utilized to create a model representing the characteristics of each class of the training sequences.

During the testing phase, we follow the same steps to extract interest points, build descriptors and assign codebook membership as those done during the training phase. Then Support Vector Machine (SVM) and K Nearest Neighbour (KNN) classifiers are adopted to classify each testing sequence to the most probable action type according to the model built in the training phase, and the correctly classified sequences against all sequences give the finial recognition accuracy. The following sub-sections elaborate each step of this framework.

3.1 Interest points detection

Interest points from a video sequence are localized not only along the spatial dimensions *x* and *y* but also the temporal dimension *t*. Currently there are three types of detection approaches: static features based on edges and limb shapes, dynamic features based on optical flow measurements and spatio-temporal features obtained from local video patches. *Dollár et al.*'s method [5] is an instance of the third type. In [22], this detector outperforms others both in accuracy and efficiency. Therefore, we adopt it for localizing interest points in our action recognition framework. This method uses separable linear filters and generally produces a high number of points. It assumes a stationary camera or a process that can account for camera motion. The response function is defined as follows:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel, applied only along the spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally. These are defined as $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$. In all cases the authors suggest $\omega = 4/\tau$, effectively making the response function R only two parameters σ and τ, corresponding roughly to the spatial and temporal scales of the detector. The response function gives strong response to periodic motion as well as the spatio-temporal corners. In general, any region with spatially distinguishing characteristics undergoing a complex motion can induce a strong response. Areas undergoing pure translational motion or without spatially distinguishing features will in general not induce a response, as a moving smoothed edge will only cause a gradual change in intensity at a given spatial location. The space-time interest points are extracted around the local maxima of the response function and could be regarded as low-level action features.

As mentioned in [5], a cuboid (spatial temporal video patch) is extracted around each interest point and it contains spatio-temporally windowed pixel values (Refer to Fig. 2). The size of cuboid is set to be of approximately six times the scale at which they were detected. The information contained in each cuboid is utilized to form a representative descriptor and moreover to build the action training model. The locality of cuboids facilitates the feature extraction which means preprocessing steps are not needed, such as foreground subtraction, figure tracking and alignment etc. Relying on each individual cuboid, we can obtain the appearance information from the cuboid itself as well as structural information from the distribution of all cuboids (interest points).

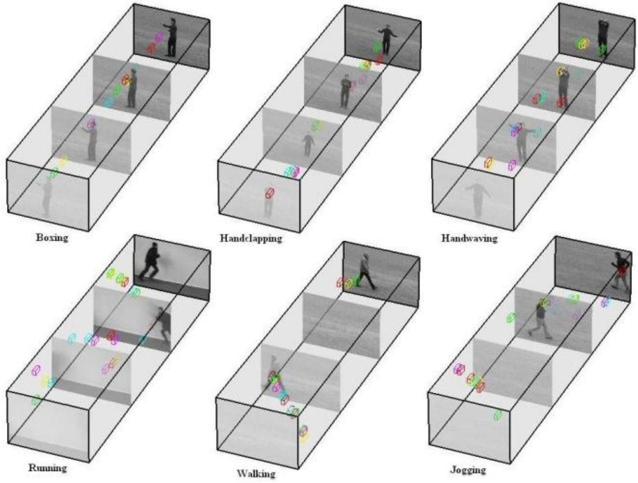


Fig. 2: Illustration of cuboids extracted around each spatial temporal interest point for each action. The cylinder indicates an action sequence, and each shows 4 frames for simplicity. Colored cuboids are displayed to express the idea that different actions have different interest points (cuboids) distribution.

3.2 Feature description

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. 3: General signal transformation process.

The general process of signal transformation is given in Fig. 3. 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. 4 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.

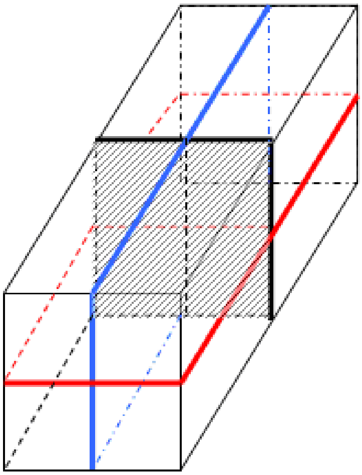


Fig. 4: Illustration of a cuboid and its internal three orthogonal planes.

Unlike the Fourier Transform, which utilizes just the sine and cosine functions, Wavelet Transforms [23] 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.

3.3 Classification methods

In the classification phase, we adopt the widely used SVM classifier. As each sequence is represented by a histogram of occurrence of each video word resulted from K-means clustering, thus the length of the histogram is equal to the number of video words, and these histograms built from all the sequences are the input to the classifier. The classifier will output the predicted class labels that the test sequences belong to.

Support Vector Machine (SVM) is a popular technique for classification. Generally, it is a binary classification algorithm that looks for an optimal hyperplane as a decision function in a high dimensional space [24].

4 Experimental Results

In this section, the experimental results of the Wavelet transform based descriptor as well as the comparison with the state-of-the-art methods are presented. The KTH human action dataset [25] is a mostly exploited while very challenging dataset. This dataset contains 600 video sequences and consists of six classes of actions (*Boxing, Handclapping, Handwaving, Running, Walking and Jogging*) performed by 25 subjects in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In order to achieve reliable accuracies, all the results presented are averaged by 20 runs.

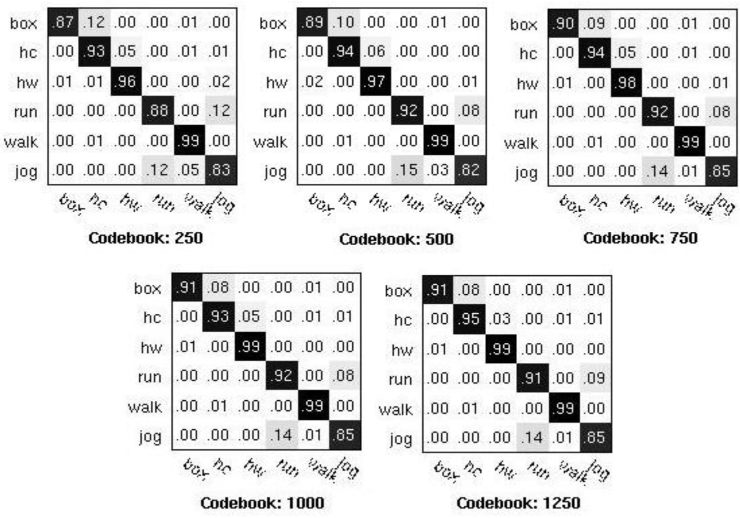
4.1 Parameter settings

In the detection phase, *Dollár et al.'s* detection method is adopted. The parameters to be set are the σ for the 2D Gaussian smoothing kernel and τ for the quadrature pair of 1D Gabor filters applied temporally. We set σ= 2.5 and τ =1.5 for all the experiments using this extraction method.

In the feature detection step, the cuboids are extracted around interest points, which are the carriers for all the transformation operations as well as gradient calculation. For feature description, we apply Daubechies wavelet transform on three orthogonal planes. Thanks to its low redundancy, the number of coefficients is only 288.

Finally, in the classification phase, we adopt non-linear Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel and the library libSVM [26] to do the classification. The best parameters *C* and *log(γ)* are selected by 5-fold cross validation in a grid search on the training data.

4.2 Result comparison



Codebook: 250

	box	hc	hw	run	walk	jog
box	.87	.12	.00	.00	.01	.00
hc	.00	.93	.05	.00	.01	.01
hw	.01	.01	.96	.00	.00	.02
run	.00	.00	.00	.88	.00	.12
walk	.00	.01	.00	.00	.99	.00
jog	.00	.00	.00	.12	.05	.83

Codebook: 500

	box	hc	hw	run	walk	jog
box	.89	.10	.00	.00	.01	.00
hc	.00	.94	.06	.00	.00	.00
hw	.02	.00	.97	.00	.00	.01
run	.00	.00	.00	.92	.00	.08
walk	.00	.01	.00	.00	.99	.00
jog	.00	.00	.00	.15	.03	.82

Codebook: 750

	box	hc	hw	run	walk	jog
box	.90	.09	.00	.00	.01	.00
hc	.00	.94	.05	.00	.01	.00
hw	.01	.00	.98	.00	.00	.01
run	.00	.00	.00	.92	.00	.08
walk	.00	.01	.00	.00	.99	.00
jog	.00	.00	.00	.14	.01	.85

Codebook: 1000

	box	hc	hw	run	walk	jog
box	.91	.08	.00	.00	.01	.00
hc	.00	.93	.05	.00	.01	.01
hw	.01	.00	.99	.00	.00	.00
run	.00	.00	.00	.92	.00	.08
walk	.00	.01	.00	.00	.99	.00
jog	.00	.00	.00	.14	.01	.85

Codebook: 1250

	box	hc	hw	run	walk	jog
box	.91	.08	.00	.00	.01	.00
hc	.00	.95	.03	.00	.01	.01
hw	.01	.00	.99	.00	.00	.00
run	.00	.00	.00	.91	.00	.09
walk	.00	.01	.00	.00	.99	.00
jog	.00	.00	.00	.14	.01	.85

Fig. 5: Confusion matrices of the Wavelet descriptor using different codebook size.

The recognition accuracy of the Wavelet based descriptor is depicted in Fig. 5 under different codebook's sizes. The accuracy for each action grows gradually and finally reaches its highest value when the codebook's size is 1250. It can be observed that the Wavelet based descriptor is discriminative for similar actions, such as *'running'* and *'jogging'*.

We compare our method to the state-of-the-art on the KTH dataset. As mentioned above, The KTH dataset is divided into the training set (16 people) and the testing set (9 people) which is the same as the experimental setup of other methods in comparison. The accuracy is averaged by 20 runs for each experiment. Table 1 shows the average accuracies of our proposed method based on Wavelet transform and methods reported by other researchers. Comparing to the existing results, our method gives the best accuracy, and outperforms the state-of-the-art in the same settings.

Table 1: Accuracy comparison on the KTH dataset.

Method	Niebles et al. [6]	Wong et al. [27]	Laptev et al. [28]	Ryoo et al. [29]	Proposed
Accuracy	83.33%	86.7%	91.8%	91.1%	93.89%

5 Conclusions

The Wavelet Transform based description method is proposed for the application of human action recognition in this paper. 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.

In future work, we will investigate other transformation techniques for video representation and recognition. The fusion of appearance information extracted from cuboids and the structural properties of each action can be another means for improving discriminative capability of local descriptors.

References

[1] 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.

[2] R. Ohbuchi, T. Takei. Shape-similarity comparison of 3D models using alpha sh-apes. In the proceedings of the Pacific Graphics. pp. 293-302, 2003.

[3] S. Savarese, J. Winn and A. Criminisi. Discriminative object class models of appearan-ce and shape by correlatons. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2, pp. 2033-2040, 2006.

[4] 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.

[5] 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.

[6] 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.

[7] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In Proceedings of IEEE International Conference on Computer Vision. pp. 726–733, 2003.

[8] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In Proceedings of IEEE International Conference on Computer Vision. Vol. 1, pp. 150-157, 2005.

[9] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. IEEE Transactionson Pattern Analysis and Machine Intelligence. Vol. 25, pp. 814 - 827, 2003.

[10] I. Laptev. On space-time interest points. International Journal of Computer Vision.Vol. 63 (2-3), pp. 107-123, 2005.

[11] Y. Ke, R.Sukthankar, M. Hebert. Efficient visual event detection using volumetric features. In Procedings of International Conference on Computer Vision. pp. 166–173, 2005.

[12] E. O. Brigham, C. K. Yuen. The fast Fourier transform. IEEE Transactions on Systems, Man and Cybernetics. Vol. 8, pp. 146-146, 1978.

[13] A. B.Watson. Image compression using the Discrete Cosine Transform. Mathema-tica Journal. Vol.4 (1), pp. 81-88, 1994.

[14] W. B. Pennebaker and J. L. Mitchell. JPEG – still image data compression standard. International Thomsan Publishing, New York. pp. 29-38, 1993.

[15] J. R. Smith, S. F. Chang. Transform features for texture classification and discrimination in large image databases. Proceedings of IEEE International. Conference on Image Processing. Vol. 3, pp. 407-411, 1994.

[16] S. Climer, S. K. Bhatia. Image database indexing using JPEG coefficients. The Journal of Pattern Recognition. Vol. 35(11), pp. 2479-2488, 2002.

[17] N. Sebe, M. S. Lew. Comparing salient point detectors. Pattern Recognition Letters. Vol. 24 (1–3), pp. 89–96, 2003.

[18] H. Schneiderman, T. Kanade. A statistical method for 3D object detection applied to faces and cars. IEEE Conference on Computer Vision and Pattern Recognition. Vol. 1, pp. 746-751, 2002.

[19] M. Gonzalez-Audicana, J. L. Saleta, R. G. Catalan, R. Garcia. Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Transactions on Geoscience and Remote Sensing. Vol. 42, pp. 1291-1299, 2004.

[20] T. H. Koornwinder. Wavelets: an elementary treatment of theory and applications. World Scientific, Signapore. pp. 49-76, 1995.

[21] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies. Image coding using wavelet transform. IEEE Transactions on Image Processing. Vol. 1, pp. 205-220.1992.

[22] L. Shao and R. Mattivi. Feature Detector and Descriptor Evaluation in Human Action Recognition. Proceedings of ACM International Conference on Image and Video Retrieval (CIVR), Xi'an, China, July 2010.

[23] M. A. Cody. The fast wavelet transform beyonds Fourier transforms. Dr. Dobb's Journal, 1992.

[24] N. Cristianini and J. Shawe-Taylor, An introduction to Support Vector Machines. Cam-bridge University Press, Cambridge. 2000.

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, pages 32-36, Vol.3, August 2004.

[26] C. C. Chang, C. J. Lin. LIBSVM: a library for support vector machines, 2001.

Software available at : http://www.csie.ntu.edu.tw/~cjlin

[27] S. F. Wong, R. Cipolla. Extracting spatiotemporal interest points using global info-rmation. In Proceedings of IEEE International Conference on Computer Vision. pp. 1-8, 2007.

[28] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld. Learning realistic human actions from movies. IEEE Conference on Computer Vision and Pattern Recognition. pp.1-8, 2008.

[29] M. S. Ryoo and J. K Aggarwal. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. International Conference on Computer Vision (ICCV), Kyoto, Japan, October 2009.