Human Action Recognition Using Non-separable Oriented 3D Dual-Tree Complex Wavelets

Rashid Minhas[1], Aryaz Baradarani[1], Sepideh Seifzadeh[2], and Q.M. Jonathan Wu[1]

[1] Department of Electrical and Computer Engineering
minhasr@uwindsor.ca, aryaz@ieee.org
[2] School of Computer Science
University of Windsor, Ontario, N9B3P4, Canada
{seifzad,jwu}@uwindsor.ca
http://www.uwindsor.ca/cvss

Abstract. This paper introduces an efficient technique for simultaneous processing of video frames to extract spatio-temporal features for fine activity detection and localization. Such features, obtained through motion-selectivity attribute of 3D dual-tree complex wavelet transform (3D-DTCWT), are used to train a classifier for categorization of an incoming video. The proposed learning model offers three core advantages: 1) significantly faster training stage than traditional supervised approaches, 2) volumetric processing of video data due to the use of 3D transform, 3) rich representation of human actions in view of directionality and shift-invariance of DTCWT. No assumptions of scene background, location, objects of interest, or point of view information are made for activity learning whereas bidirectional 2D-PCA is employed to preserve structure and correlation amongst neighborhood pixels of a video frame. Experimental results compare favorably to recently published results in literature.

1 Introduction

In recent years, research community has witnessed considerable interest in activity recognition due to its imperative applications in different areas such as human-computer interface, gesture recognition, video indexing and browsing, analysis of sports events and video surveillance. Despite the fact that initially good results were achieved, traditional action recognition approaches have inherent limitations. Different representations have been proposed in action recognition such as optical flow [3], geometrical modeling of local parts space-time templates, and hidden Markov model (HMM) [4]. Geometrical model [1,8,12] of local human parts are used to recognize the action using static stances in a video sequence which match a sought action. In space-time manifestation [21], outline of an object of interest is characterized in space and time using silhouette. The volumetric analysis of video frames has also been proposed [6] where video alignment is usually unnecessary and space-time features contain descriptive information for an action classification. In [6] promising results are achieved

H. Zha, R.-i. Taniguchi, and S. Maybank (Eds.): ACCV 2009, Part III, LNCS 5996, pp. 226–235, 2010.

assuming that background is known for preliminary segmentation. For action recognition use of space-time interest points has been proved to be a thriving technique [11] without any requirement for pre-segmentation or tracking of individual dynamic objects in a video. Recently, classifiers that embed temporal information in a recognition process [18] and event recognition in still images [10] have been proposed where distinct scene and object recognition schemes are applied to identify an event.

A novel action learning framework using 3D dual-tree complex wavelet transform (3D-DTCWT) is proposed to process volumetric data of a video sequence instead of searching a specific action through feature detection in individual frames and finding their temporal behavior. Proposed by Kingsbury [9], 2D dual-tree complex wavelet transform has two important properties; the transformation is nearly shift-invariant and has a good directionality in its subbands. The idea of multiresolution transform for motion analysis was proposed in [5] and further developed as 3D wavelet transform in video denoising by Selesnick et al. [15], which is an important step to overcome the drawbacks of previously introduced video denoising techniques.These limitations are due to separable implementation of 1D transforms in a 3D space, and also due to an artifact called *checkerboard* effect which has been extensively explained in an excellent survey on theory, design and application of DTCWT in [17]. Selesnick et al. refined their work in [16] introducing non-separable 3D wavelet transform using Kingsbury's filter banks [9,17] to provide an efficient representation of *motion*-selectivity.

For real time processing, complex wavelet coefficients of different subbands are represented by lower dimension feature vectors obtained using bidirectional 2D-PCA, i.e., a variant of 2D-PCA [20]. Extreme learning machine (ELM) is applied to classify the actions represented by feature vectors. ELM is a supervised learning framework [7], single hidden layer feedforward neural network, that is trained at speed of thousands times faster than traditional learning schemes such as gradient descent approach in traditional neural networks.

2 Dual-Tree Complex Filter Banks

Consider the two-channel dual-tree filter bank implementation of the complex wavelet transform. Shown in Fig. 1, the primal filter bank in each level defines the real part of the wavelet transform. The dual filter bank depicts the imaginary part when both the primal and dual filter banks work in parallel to make a dual-tree structure. Recall that the scaling and wavelet functions associated with the analysis side of the primal are defined by two-scale equations $\phi_h(t) = 2\sum_n h_0[n]\phi_h(2t-n)$ and $\psi_h(t) = 2\sum_n h_1[n]\phi_h(2t-n)$. The scaling function ϕ_f and wavelet function ψ_f in the synthesis side of the primal are similarly defined via f_0 and f_1. The same is true for the scaling functions ($\widetilde{\phi}_h$ and $\widetilde{\phi}_f$) and wavelet functions ($\widetilde{\psi}_h$ and $\widetilde{\psi}_f$) of the dual filter bank $\widetilde{\mathbf{B}}$. The dual-tree filter bank defines analytic complex wavelets $\psi_h + j\widetilde{\psi}_h$ and $\widetilde{\psi}_f + j\psi_f$, if the wavelet functions of the two filter banks form Hilbert transform pairs. Specifically, the analysis wavelet $\widetilde{\psi}_h(t)$ of $\widetilde{\mathbf{B}}$ is the Hilbert transform of the analysis wavelet

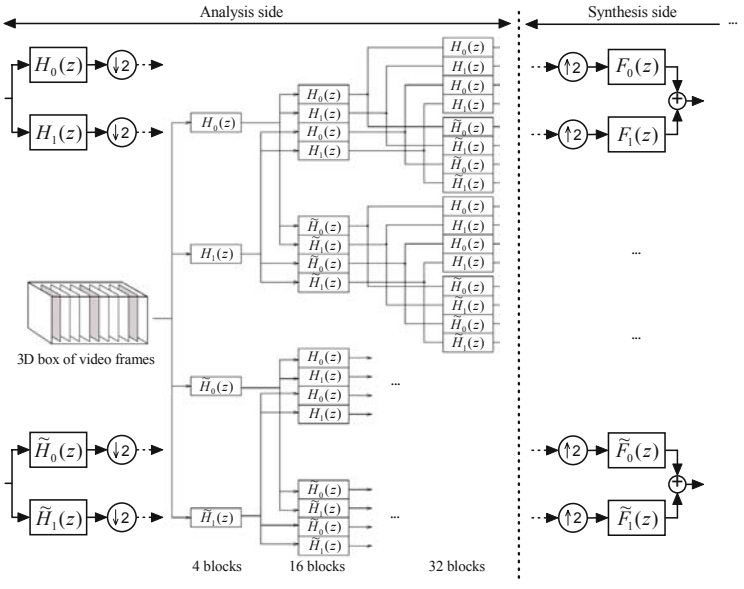


Fig. 1. Typical schematic of a 3D-DTCWT structure with the real and imaginary parts of a complex wavelet transform. Only the analysis side is shown in details in this figure.

$\psi_h(t)$ of $\mathbf{B}$, and the synthesis wavelet $\psi_f(t)$ of $\mathbf{B}$ is the Hilbert transform of $\widetilde{\psi}_f(t)$. That is, $\widetilde{\Psi}_h(\omega) = -j\mathtt{sign}(\omega)\Psi_h(\omega)$ and $\Psi_f(\omega) = -j\mathtt{sign}(\omega)\widetilde{\Psi}_f(\omega)$, where $\Psi_h(\omega)$, $\Psi_f(\omega)$, $\widetilde{\Psi}_h(\omega)$, and $\widetilde{\Psi}_f(\omega)$ are the Fourier transforms of wavelet functions $\psi_h(t)$, $\psi_f(t)$, $\widetilde{\psi}_h(t)$, and $\widetilde{\psi}_f(t)$ respectively, `sign` represents the signum function, and j is the square root of -1. This introduces limited redundancy and allows the transform to provide approximate shift-invariance and more directionality selection of filters [9,17] while preserving properties of a perfect reconstruction and computational efficiency with improved frequency responses. It should be noted that these properties are missing in discrete wavelet transform (DWT). The filter bank $\mathbf{B}$ constitutes a biorthogonal filter bank [19] if and only if its filters satisfy the no-distortion condition

$$H_0(\omega)F_0(\omega) + H_1(\omega)F_1(\omega) = 1 \tag{1}$$

and the no-aliasing condition

$$H_0(\omega + \pi)F_0(\omega) + H_1(\omega + \pi)F_1(\omega) = 0. \tag{2}$$

The above no-aliasing condition is automatically satisfied if $H_1(z) = F_0(-z)$ and $F_1(z) = -H_0(-z)$. The wavelets of the dual filter bank exhibits similar characteristics, i.e., $\widetilde{H}_1(z) = \widetilde{F}_0(-z)$ and $\widetilde{F}_1(z) = -\widetilde{H}_0(-z)$ where z refers to the z-transform. A complete procedure of DTCWT design and its application for moving object segmentation is presented in [22] and [2] respectively.

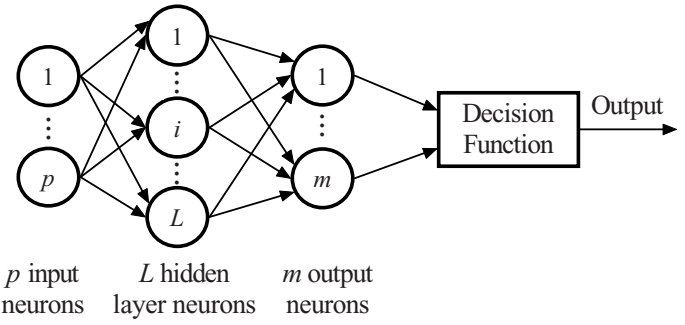


Fig. 2. Simplified structure of ELM

2.1 Non-separable 3D Dual-Tree Complex Wavelet Transform

Generally, wavelet bases are optimal for the category of one dimensional signals. In case of 2D (two dimensional), however, the scalar 2D discrete wavelet transform (2D-DWT) cannot be an optimal choice [17] because of the weak line (curve)-singularities of DWT although it is still better than the discrete cosine transform (DCT). In video, however, the situation is even worse and the edges of objects move in more spatial directions (motion) yielding a 3D edge effect. The 3D-DTCWT includes a number of wavelets which are expansive than real 3D dual-tree wavelet transform. This is related to the real and imaginary parts of a 3D complex wavelet with two wavelets in each direction. Fig. 1 shows the structure of a typical 3D-DTCWT. Note that the wavelets associated with 3D-DTCWT are free of the checkerboard effect. The effect remains disruptive for both the separable 3D-CWT (complex wavelet transform) and 3D-DWT. Recall that for 3D-DTCWT, in stage three (the third level of the tree), there are 32 subbands from which 28 are counted as wavelets excluding the scaling subbands, compared with the 7 wavelets for separable 3D transforms. Thus, 3D-DTCWT can better localize motion in its several checkerboard-free directional subbands compared with 2D-DWT and separable 3D-DWT with less number of subbands and checkerboard phenomena. It should be noted that there is a slight abuse of using the term subband here. It is more reasonable to use the terms of 'blocks' or 'boxes' instead of 'subbands' in a 3D wavelet structure.

2.2 Extreme Learning Machine

It is a well known fact that the slow learning speed of feedforward neural networks (FNN) has been a major bottleneck in different applications. Huang et al. [7] showed that single-hidden layer feedforward neural network, also termed as extreme learning machine (ELM), can exactly learn N distinct observations for almost any nonlinear activation function with at most N hidden nodes (see Fig. 2). Unlike the popular thinking that network parameters need to be tuned, one may not adjust the input weights and first hidden layer biases but they are randomly assigned. Such an approach has been proven to perform learning at an extremely fast speed, and obtains better generalization performance

for activation functions that are infinitely differentiable in hidden layers. For N arbitrary distinct samples (x_i, γ_i) where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]' \in R^p$ and $\gamma_i = [\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{im}]' \in R^m$ (The superscript " $'$ " represents the transpose) represent input samples and label respectively. A standard ELM with L hidden nodes and an activation function $g(x)$ is modeled as

$$\sum_{i=1}^{L} \beta_i g(x_l) = \sum_{i=1}^{L} \beta_i g(w_i.x_l + b_i) = o_l,\ l \in \{1, 2, 3, \ldots, N\} \tag{3}$$

where $w_i = [w_{i1}, w_{i2}, \ldots, w_{ip}]'$ and $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}]'$ represent the weight vectors connecting the input nodes to an ith hidden node and from the ith hidden node to the output nodes respectively. b_i shows a threshold for an ith hidden node, and $w_i.x_l$ represents inner product of w_i and x_l. The above modeled ELM can reliably approximate N samples with zero error as

$$\sum_{l=1}^{L} \|o_l - \gamma_l\| = 0 \tag{4}$$

$$\sum_{i=1}^{L} \beta_i g(w_i.x_l + b_i) = \gamma_l,\ l \in \{1, 2, \ldots, N\} \tag{5}$$

The above N equations can be written as $\Upsilon\beta = \Gamma$ where $\beta = [\beta_1', \ldots, \beta_L']'_{L\times m}$ and $\Gamma = [\gamma_1', \ldots, \gamma_N']'_{N\times m}$. In this formulation Υ is called the hidden layer output matrix of ELM where ith column of Υ is the output of ith hidden node output with respect to inputs $x_1, x_2, \ldots, x_N$. If the activation function g is infinitely differentiable, the number of hidden nodes are such that $L \ll N$. Thus

$$\Upsilon = (w_1, \ldots, w_L, b_1, \ldots, b_L, x_1, \ldots, x_N) \tag{6}$$

Traditionally, training of ELM requires minimization of an error function ε in terms of the defined parameters as

$$\varepsilon = \textstyle\sum_{l=1}^{N} (\sum_{i=1}^{L} \beta_i g(w_i x_l + b_i) - \gamma_l)^2 \tag{7}$$

3 Proposed Algorithm

Our proposed framework assigns an action label to an incoming video based upon observed activity. Initial feature vectors are extracted after segmenting moving objects in various frames of an input sequence. Each video frame is converted to gray level and a square dimension matrix. No further pre-processing is applied to input data and we assume no additional information about location, view point, activity, background and data acquisition constraints. We do not require knowledge of background or static objects since a robust moving object detection and segmentation algorithm is applied to extract movements in a video [2]. After segmentation, video frames contain moving object information only. Applying

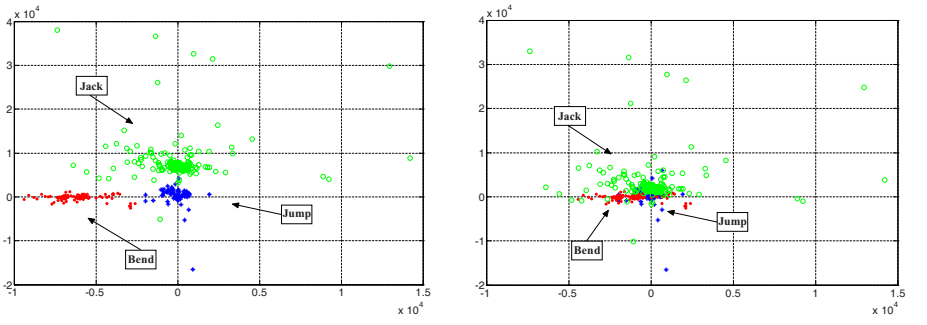


Fig. 3. Spatio-temporal information captured by (a) bidirectional 2D-PCA, (b) PCA

3D-DTCWT on segmented video sequence of size (Q, M, P) results into a box of video frames of size $(Q/2, M/2, P/2)$ where Q, M, and P represent rows, columns and number of frames respectively.

To avoid the curse of dimensionality bidirectional 2D-PCA is employed which requires multiplication between two feature matrices computed in row and column directions using 2D-PCA [20]. As opposed to PCA, 2D-PCA is based on 2D image matrices rather than 1D vectors, therefore the image matrix does not need to be vectorized prior to feature extraction. We propose a modified scheme to extract features using 2D-PCA by computing two image covariance matrices of the square training samples in their original and transposed forms respectively while the training image mean need not be necessarily equal to zero. The vectorization of mutual product of such 2D-PCA matrices results into a considerably smaller sized *subband* feature vectors that retain better structural and correlation information amongst neighboring pixels. Fig. 4 shows better ability of bidirectional 2D-PCA to represent the spatio-temporal information of various actions performed by actor Daria. Fig. 4(a) and (b) are plotted against three different videos that contain activity of Jack, Bend and Jump respectively. The first two components of subband feature vectors obtained using bidirectional 2D-PCA and traditional PCA are plotted. In Fig. 4(a), the separability of different action classes is noticeable whereas components are merged for the feature vectors obtained using PCA (Fig. 4(b)). Procedure of the proposed algorithm is briefly summarized below.

`INPUT:` A video sequence consisting of frames $a_i, 1 \leq i \leq P$

`OUTPUT:` Recognized Action

`Step 1:` Apply segmentation and extract video frames $b_i, 1 \leq i \leq P$ with moving objects only (see Fig. 4).

`Step 2:` Compute coefficients $C_f^d, d \leq 28, f \leq P/2$ (d, f represent filters orientation and number of coefficient matrices) using 3D-DTCWT on volumetric video data $b_i, 1 \leq i \leq P$.

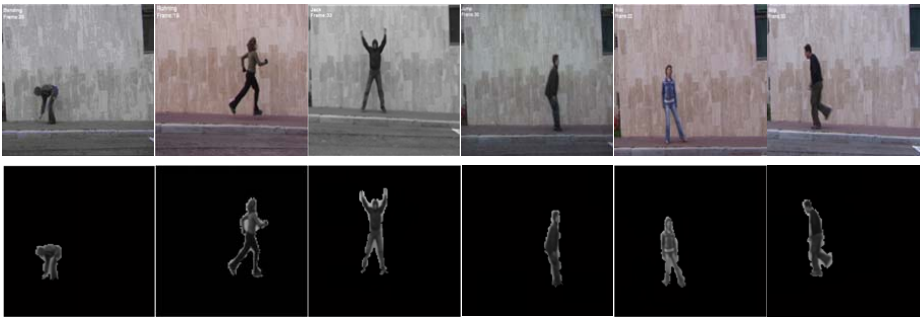
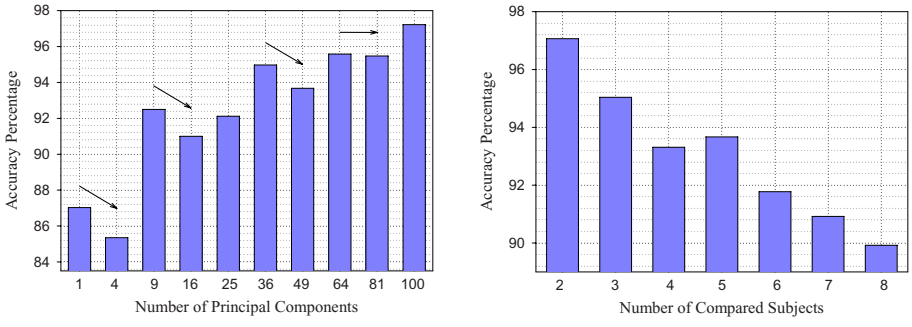


Fig. 4. First row: Randomly selected actions. Bending, Running, Jack, Jump, Side, and Skip (left to right). Second row: Segmented frames using our technique in [2].



(a) (b)

Fig. 5. Accuracy analysis. (a) Video classification for increasing size of feature vectors, (b) Varying number of compared videos.

`Step 3:` Compute feature vectors using coefficients computed in Step 2. Calculate coefficient covariance matrices G:
$G_d = \frac{1}{f}\sum_{i=1}^{f}(C_i^d - \overline{C})'(C_i^d - \overline{C})$, $\overline{C} = \sum_{i=1}^{f} C_i^d$, $d \in \{1, 2, \ldots, 28\}, i \leq f$.

`Step 4:` Evaluate the maximizing criteria $J(X)$ for projection vector X. $J(X) = X'G_dX, d \in \{1, 2, \ldots, 28\}$.

`Step 5:` 2D-PCA for coefficient matrix C_i^l is represented as $Y_d = C_i^d X, d \in \{1, 2, \ldots, 28\}, i \leq f$.

`Step 6:` Determine bidirectional 2D-PCA as $Y_d^{bi} = Y_d'.Y_d, d \in \{1, 2, \ldots, 28\}$.

`Step 7:` Convert matrices $Y_d^{bi}, d \in \{1, 2, \ldots, 28\}$ into vector form.

`Step 8:` Train ELM using training feature vectors (computed in Steps 1–7).

`Step 9:` Test the action in an incoming video using feature vectors.

4 Results and Discussion

To test the performance of our proposed method, publicly available datasets [6] are used in experiments. These datasets contain different subjects which perform

Table 1. Confusion table for dataset [6]

	Bend	Jump	Jack	Side	Walk	Run	Skip	Wave1	Wave2
Bend	9								
Jump		8		1					
Jack			9						
Side			1	8					
Walk					9				
Run						8			1
Skip					8		1		
Wave1								9	
Wave2									9

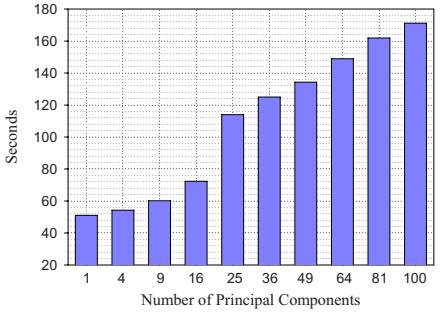


Fig. 6. Computational complexity of our proposed method

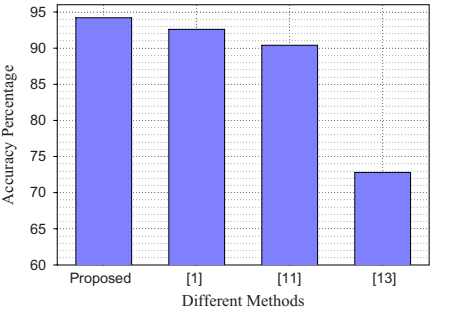


Fig. 7. Performance analysis of different methods

nine distinct actions and therefore we have various videos with varying number of frames. In Fig. 4, first row shows some randomly selected video frames for various subjects along with their segmentation in second row. A leave-one-out cross validation scheme is applied whereas results presented in this section are averaged values for 10 runs of the same experiment through random selection of subjects and/or actions in the dataset.

It is also worth to point out that the dimension of individual feature vectors may also affect the video classification since larger feature vectors retain more information at the expense of higher computational complexity. However, anticipating improved classification by monotonically increasing the size of feature vectors is not a rationale approach. As presented in Fig. 5(a), the accuracy is not constantly increasing by raising the dimensionality of feature vectors; especially classification precision is dwindling or remains constant at arrow locations while the computational complexity of classification is ever-increasing. In our experiments, we achieve reasonable degree of action recognition for the size of feature vectors varying from 36 to 100. In the past, as per the best knowledge of authors, classification accuracy has been reported for a fixed number of training and testing actions/subjects whereas it is an interesting investigation

to judge the accuracy of a classifier by analyzing its performance for randomly selected combinations of training and testing videos. As shown in Fig. 5 (b), our proposed method achieves an average classification precision of 94.2% for varying number and combinations of subjects/videos. Insightful investigation reveals the fact that for a larger number of compared videos of the same or different actions/object has higher probability for false alarms [see Fig. 5(b)] due to apparently the same activity observed in small number of adjacent frames such as standing, similar movements in between repetition of an action etc. Fig. 6 represents the computational complexity (combined for training and testing) of our proposed classifier for activity recognition and classification after extraction of feature vectors of all videos of the dataset [6] with respect to the employed processor. Categorization time is another important factor to measure the performance of a classification framework. Our proposed method completes classification of *subband* feature vectors at considerably faster speed. This is an additional improvement on existing methods since proposed technique requires significantly less time for identification and categorization due to enormously simple structure of a classifier comprising of only one hidden layer of neurons yet producing promising results. Table 1 shows confusion table, with achieved accuracy of 95.04%, for a random combination of one and three videos used for testing and training purpose respectively.

Fig. 7 presents accuracy analysis of various methods for human action recognition; it is clear that our proposed scheme outperforms well-known techniques in terms of accuracy and computational complexity. We achieve 94.2% correct action labeling that is an average value for varying number and combinations of randomly selected videos instead of relying on a specific number and/or combination of video sets.

5 Conclusion

A new human action recognition framework is presented based on volumetric video data processing rather than frame by frame analysis. Our method assumes no *a priori* knowledge about activity, background, view points and/or acquisition constraints in an arriving video. Shift-invariance and motion selectivity properties of 3D-DTCWT support reduced artifacts and resourceful processing of a video for better quality and well-localized activity categorization. *Subband* feature vectors, computed using bidirectional 2D-PCA, are input to an ELM that offers classification at considerably higher speed in comparison with other learning approaches such as classical neural networks, SVM and AdaBoost.

Acknowledgment

This work was partially funded by the Natural Science and Engineering Research Council of Canada.

References

1. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. In: Proc. of Int. Conf. on CV (2007)
2. Baradarani, A., Wu, J.: Moving object segmentation using the 9/7–10/8 dual-tree complex filter bank. In: Proc. of the 19th IEEE Int. Conf. on PR, Florida, pp. 7–11 (2008)
3. Black, M.J.: Explaining optical flow events with parameterized spatio-temporal models. In: Proc. of Int. Conf. on CVPR, pp. 1326–1332 (1999)
4. Brand, M., Oliver, N., Pentland, A.: Coupled HMM for Complex Action Recognition. In: Proc. of Int. Conf. on CVPR (1997)
5. Burns, T.J.: A non-homogeneous wavelet multiresolution analysis and its application to the analysis of motion, PhD thesis, Air Force Institute of Tech. (1993)
6. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space time shapes. IEEE Trans. on PAMI, 2247–2253 (2007)
7. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing (2005)
8. Jiang, H., Drew, M.S., Li, Z.N.: Successive convex matching for action detection. In: Proc. of Int. Conf. on CVPR (2006)
9. Kingsbury, N.G.: Complex wavelets for shift invariant analysis and filtering of signals. Journal of Applied and Computational Harmonic Analysis 10(3), 234–253 (2001)
10. Li, L.-J., Li, F.-F.: What, where and who? Classifying events by scene and object recognition. In: Proc. of Int. Conf. on CV (2007)
11. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: Proc. of Int. Conf. on CVPR (2008)
12. Mori, G., Ren, X., Efros, A.A., Malik, J.: Recovering human body configurations: combining segmentation and recognition. In: Proc. of Int. Conf. on CVPR (2004)
13. Niebels, J. L.F.-F.: A hierarchical model of shape and appearance for human action classification. In: Proc. of Int. Conf. on CVPR (2007)
14. Selesnick, I.W.: Hilbert transform pairs of wavelet bases. IEEE Signal Processing Letters 8, 170–173 (2001)
15. Selesnick, I.W., Li, K.Y.: Video denoising using 2D and 3D dual-tree complex wavelet transforms, Wavelet Applications in Signal and Image. In: Proc. SPIE 5207, San Diego (August 2003)
16. Selesnick, I.W., Shi, F.: Video denoising using oriented complex wavelet transforms. In: Proc. of the IEEE Int. Conf. on Acoust., Speech, and Signal Proc., May 2004, vol. 2, pp. 949–952 (2004)
17. Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.G.: The dual-tree complex wavelet transform – a coherent framework for multiscale signal and image processing. IEEE Signal Processing Magazine 6, 123–151 (2005)
18. Smith, P., Victoria, N.D., Shah, M.: TemporalBoost for event recognition. In: Proc. of Int. Conf. on CV (2005)
19. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley, Cambridge (1996)
20. Yang, J., Zhang, D., Frangi, F., Yang, J.-Y.: Two-dimensional PCA: a new approach to appearance based face representation and recognition. IEEE Trans. on PAMI (1), 131–137 (2004)
21. Yilmaz, A., Shah, M.: Actions sketch: a novel action representation. In: Proc. of Int. Conf. on CVPR (2005)
22. Yu, R., Baradarani, A.: Sampled-data design of FIR dual filter banks for dual-tree complex wavelet transforms. IEEE Trans. on Signal Proc. 56(7), 3369–3375 (2008)