Neurocomputing 73 (2010) 1906–1917

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Human action recognition using extreme learning machine based on visual vocabularies

Rashid Minhas [a], Aryaz Baradarani [a], Sepideh Seifzadeh [b], Q.M. Jonathan Wu [a,*]

[a] Department of Electrical and Computer Engineering, University of Windsor, Ontario, Canada N9B 3P4
[b] School of Computer Science, University of Windsor, Ontario, Canada N9B 3P4

ARTICLE INFO

Available online 18 March 2010

Keywords:
Extreme learning machine
Activity recognition
3D dual-tree complex wavelet transform
Two-dimensional PCA
Video classification

ABSTRACT

This paper introduces a novel recognition framework for human actions using hybrid features. The hybrid features consist of *spatio-temporal* and *local static* features extracted using motion-selectivity attribute of 3D dual-tree complex wavelet transform (3D DT-CWT) and affine SIFT local image detector, respectively. The proposed model offers two core advantages: (1) the framework is significantly faster than traditional approaches due to volumetric processing of images as a '3D box of data' instead of a frame by frame analysis, (2) rich representation of human actions in terms of reduction in artifacts in view of the promising properties of our recently designed full symmetry complex filter banks with better directionality and shift-invariance properties. No assumptions about scene background, location, objects of interest, or point of view information are made whereas bidirectional two-dimensional PCA (2D-PCA) is employed for dimensionality reduction which offers enhanced capabilities to preserve structure and correlation amongst neighborhood pixels of a video frame.

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. Traditional activity recognition approaches have inherent limitations for robust categorization and localization of actions in presence of non-stationary background, varying posture and size of moving objects, and occlusions. Whereas variations in illumination, behavior and mutual interactions of dynamic objects in scenes add further complications to recognition tasks.

1.1. Related work

For action recognition, different representations have been proposed such as optical flow [2], geometrical modeling of local parts space-time templates, and hidden Markov model (HMM) [3] (however, large number of features may result in higher computational load). Generally, the precision of optical flow estimation is reliant upon tribulations in aperture and properties of the surface being captured. In [1,8,15], geometrical models 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, outline of an object of interest is characterized in space and time using silhouette or body contour to model an action [5,8,15,25,28,36]. 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 assuming that background is known for preliminary segmentation. For action recognition, the use of space-time interest points has been proved to be a thriving technique [11,27,29,30,32,35] without any requirement for pre-segmentation or tracking of individual dynamic objects in a video. To improve classification performance, both shape and spatio-temporal features are combined [12,13,17]; some researchers have proposed to integrate *a priori* information of a scene into recognition process which may include operations like stabilization, video trimming and segmentation using readily available masks or automated detection of movements in consecutive frames [6,28,31].

The spatio-temporal (ST) features [12,30] and space-time interest points (STIP) features [11,35] have successfully been used in action recognition. The ST feature detector produces dense set of features with a reasonable performance in activity recognition tasks. The detector applies two separate linear filters to the spatial and temporal dimensions, respectively, instead of using only one 3D filter which requires higher computational time. The ST volumes around interest points, detected using locally maximum response of both filters, are extracted for further processing. To detect events in a video sequence,

* Corresponding author.
E-mail addresses: minhasr@uwindsor.ca (R. Minhas), baradar@uwindsor.ca (A. Baradarani), seifzad@uwindsor.ca (S. Seifzadeh), jwu@uwindsor.ca (Q.M. Jonathan Wu).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.01.020

the extraction of STIP features is based on the idea of Harris and Förstner interest point operators but extended to spatio-temporal domain by acquiring the image values in space-time which have large variations in both spatial and temporal dimensions. Moreover, STIP features can be represented using three different local space-time descriptors, i.e., Histogram of Oriented Gradients (HoG), Histogram of Optical Flow (HoF) and the combination of both termed as HnF.

Using 3D dual-tree complex wavelet transform (3D DT-CWT), in this paper, a novel action recognition framework is proposed that processes volumetric data of a video sequence instead of searching a specific action through feature detection in individual frames and finding their temporal behavior. Dual-tree complex wavelet transform (DT-CWT) is constructed by designing an appropriate pair of orthogonal or biorthogonal filter banks that work in parallel. Proposed by Kingsbury [9], 2D DT-CWT 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 [4] and further developed as 3D wavelet transform in video denoising by Selesnick et al. [19,20], which is an important step to overcome the limitations caused by the 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 DT-CWT in [21]. Selesnick et al. refined their work in [20] by introducing non-separable 3D wavelet transform using Kingsbury's filter banks [9,21] to provide an efficient representation of *motion*-selectivity (the so-called *directional*-selectivity of DT-CWT in two-dimensional space).

To determine *spatio-temporal* features, complex wavelet coefficients of different subbands are represented by lower dimension feature vectors obtained using bidirectional two-dimensional PCA (2D-PCA), i.e. a variant of 2D-PCA [23]. Bidirectional 2D-PCA performs in both row and column-wise directions and better preserves the correlation amongst neighboring pixels. To extract *local static* features, affine SIFT descriptors are computed for patches around detected interest points. A pruning strategy is applied to eliminate the descriptors which belong to static parts of a scene in a video since such information is not significant for accurate activity recognition. Finally, we construct visual vocabularies using both kinds of features to be input to a classifier to recognize an action present in an arriving video. We do not present discussion on construction of visual vocabularies in this paper; however, interested readers should refer to [13] for in-depth details. Extreme learning machine (ELM) is a supervised learning framework [7], single hidden layer feedforward neural network, that is trained at speed roughly thousand times faster than traditional learning schemes such as gradient descent approach in traditional neural networks. ELM is applied to classify the actions represented by visual vocabularies.

The rest of paper is organized as follows. In Section 2, preliminary information about dual-tree complex filter banks and learning classifier are presented. Sections 3 and 4 comprise of proposed algorithm and several illustrative examples followed by results and discussions. The paper is concluded in Section 5.

2. Preliminaries

2.1. Dual-tree complex filter banks

Consider the two-channel dual-tree filter bank implementation of the complex wavelet transform. Shown in Fig. 1(a), the primal filter bank $\mathbf{B}$ in each level defines the real part of the wavelet transform. The dual filter bank $\tilde{\mathbf{B}}$ shown in Fig. 1(b) depicts the imaginary part when both the primal and dual filter banks work in parallel to make a dual-tree structure. Recall

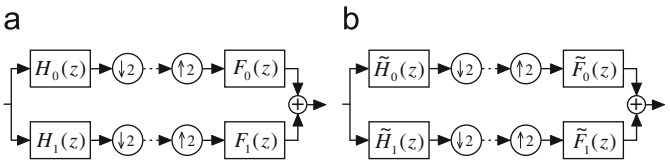

Fig. 1. (a) The primal filter bank $\mathbf{B}$ and (b) the dual filter bank $\tilde{\mathbf{B}}$.

that the scaling and wavelet functions associated with the analysis side of $\mathbf{B}$ 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 $\mathbf{B}$ are similarly defined via f_0 and f_1. The same is true for the scaling functions ($\tilde{\phi}_h$ and $\tilde{\phi}_f$) and wavelet functions ($\tilde{\psi}_h$ and $\tilde{\psi}_f$) of the dual filter bank $\tilde{\mathbf{B}}$.

The dual-tree filter bank defines analytic complex wavelets $\psi_h + j\tilde{\psi}_h$ and $\tilde{\psi}_f + j\psi_f$, if the wavelet functions of the two filter banks form Hilbert transform pairs. Specifically, the analysis wavelet $\tilde{\psi}_h(t)$ of $\tilde{\mathbf{B}}$ is the Hilbert transform of the analysis wavelet $\psi_h(t)$ of $\mathbf{B}$, and the synthesis wavelet $\psi_f(t)$ of $\mathbf{B}$ is the Hilbert transform of $\tilde{\psi}_f(t)$. That is, $\tilde{\Psi}_h(\omega) = -j\,\mathrm{sign}(\omega)\Psi_h(\omega)$ and $\Psi_f(\omega) = -j\,\mathrm{sign}(\omega)\tilde{\Psi}_f(\omega)$, where $\Psi_h(\omega)$, $\Psi_f(\omega)$, $\tilde{\Psi}_h(\omega)$, and $\tilde{\Psi}_f(\omega)$ are the Fourier transforms of wavelet functions $\psi_h(t)$, $\psi_f(t)$, $\tilde{\psi}_h(t)$, and $\tilde{\psi}_f(t)$, respectively, sign represents the signum function, and j is the square root of -1 [37]. This introduces limited redundancy and allows the transform to provide approximate shift-invariance and more directionality selection of filters [9,21] 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 [22] 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

$$H_1(z) = F_0(-z) \quad \text{and} \quad F_1(z) = -H_0(-z). \tag{3}$$

The wavelet filter banks of $\tilde{\mathbf{B}}$ exhibits similar characteristics:

$$\tilde{H}_1(z) = \tilde{F}_0(-z) \quad \text{and} \quad \tilde{F}_1(z) = -\tilde{H}_0(-z). \tag{4}$$

where z refers to the z-transform.

2.1.1. 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 [21,22] because of the weak line (curve)-singularities of DWT although its performance 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 DT-CWT 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. 2 shows the structure of a typical 3D DT-CWT. Note that the wavelets associated with 3D DT-CWT 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 DT-CWT, 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 DT-CWT can

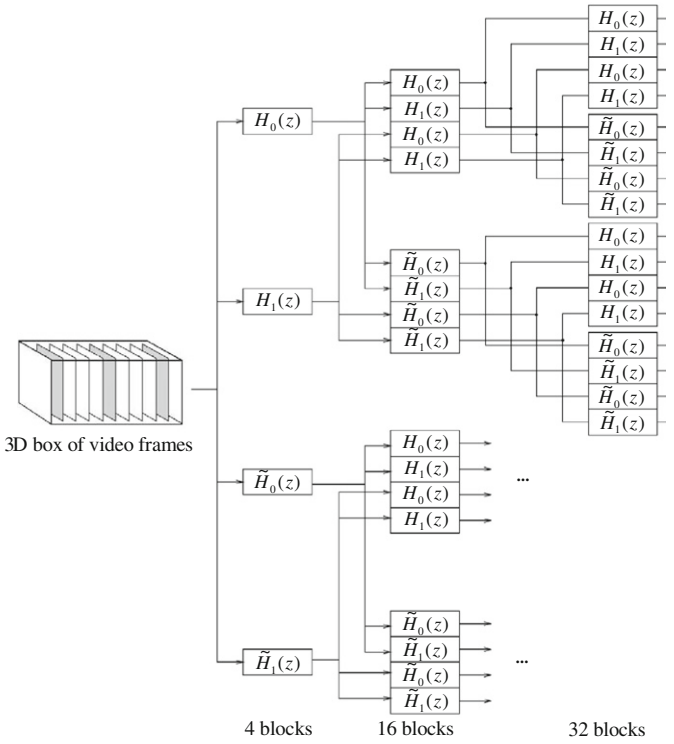

Fig. 2. Typical schematic of filters in a 3D DT-CWT structure with the real and imaginary parts of a complex wavelet transform. 28 of the 32 subbands are wavelets excluding the scaling terms. Only the analysis side is shown in this figure.

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

Feedforward neural networks (FNN) have been widely used in different areas due to their approximation capabilities for nonlinear mappings using input samples. It is a well known fact that slow learning speed of FNN has been a major bottleneck in different applications. In the past theoretical research, the input weights and hidden layer biases need to be adjusted using some parameter tuning approach such as gradient descent based methods. However, gradient descent based learning techniques are generally slow due to inappropriate learning steps with significantly large latency to converge to a local maxima. Huang et al. [7] showed that single-hidden layer feedforward neural network, also termed as ELM, can exactly learn N distinct observations for almost any nonlinear activation function with at most N hidden nodes (see Fig. 3). 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 good generalization performance for activation functions that are infinitely differentiable in hidden layers. ELM converts the learning problem into a simple linear system whose output weights can be analytically determined through a generalized inverse operation of the hidden layer weight matrices. Such a learning scheme can operate at approximately thousands of times faster speed than learning

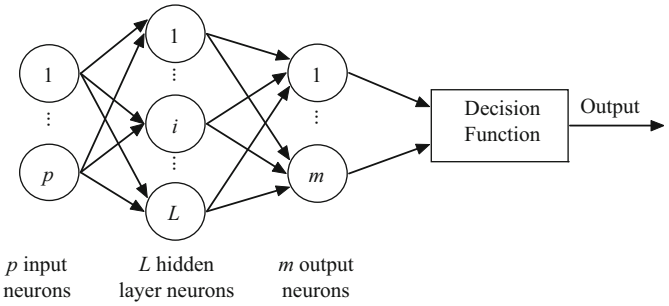

Fig. 3. Simplified structure of ELM.

strategy of traditional feedforward neural networks like back propagation (BP) algorithm [7]. Improved generalization performance with the smallest training error and the norm of weights demonstrate its superior classification capability for real-time applications at an exceptionally fast pace without any learning bottleneck. 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), a standard ELM with L hidden nodes and an activation function $g(x)$ is modeled by

$$\sum_{i=1}^{L}\beta_i g(x_l)=\sum_{i=1}^{L}\beta_i g(w_i\cdot x_l+b_i)=o_l,\quad l\in\{1,2,3,\ldots,N\},\tag{5}$$

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 i th hidden node and from the i th hidden node to the output nodes, respectively, b_i shows a threshold for an i th hidden node and w_i. x_l represents the inner product of w_i and x_l. The above modeled ELM can reliably approximate N samples with zero error as

$$\sum_{l=1}^{N}\|o_l-\gamma_l\|=0\tag{6}$$

$$\sum_{i=1}^{L}\beta_i g(w_i\cdot x_l+b_i)=\gamma_l,l\in\{1,2,\ldots,N\}.\tag{7}$$

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 i th column of Υ is the output of i th hidden node 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,

$$\Upsilon=(w_1,\ldots,w_L,b_1,\ldots,b_L,x_1,\ldots,x_N).\tag{8}$$

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

$$\varepsilon=\sum_{l=1}^{N}\left(\sum_{i=1}^{L}\beta_i g(w_i x_l+b_i)-\gamma_l\right)^2,\tag{9}$$

where it is sought to minimize the error, $\varepsilon=\|\Upsilon\beta-\Gamma\|$. Traditionally unknown Υ is determined using gradient descent based scheme and the weight vector W, which is a combination of w_i, β_i, and bias parameters b_i, is tuned iteratively by

$$w_k=w_{k-1}-\rho\frac{\partial\varepsilon(W)}{\partial W}.\tag{10}$$

The learning rate ρ significantly affects the accuracy and learning speed; a small value of ρ causes the learning algorithm to converge at a significantly slower rate whereas a larger learning step leads to instability and divergence. Huang et al. [7] proposed minimum norm least-square solution for ELM to avoid aforementioned limitations encountered in conventional learning paradigm which states that the input weights and the hidden layer biases

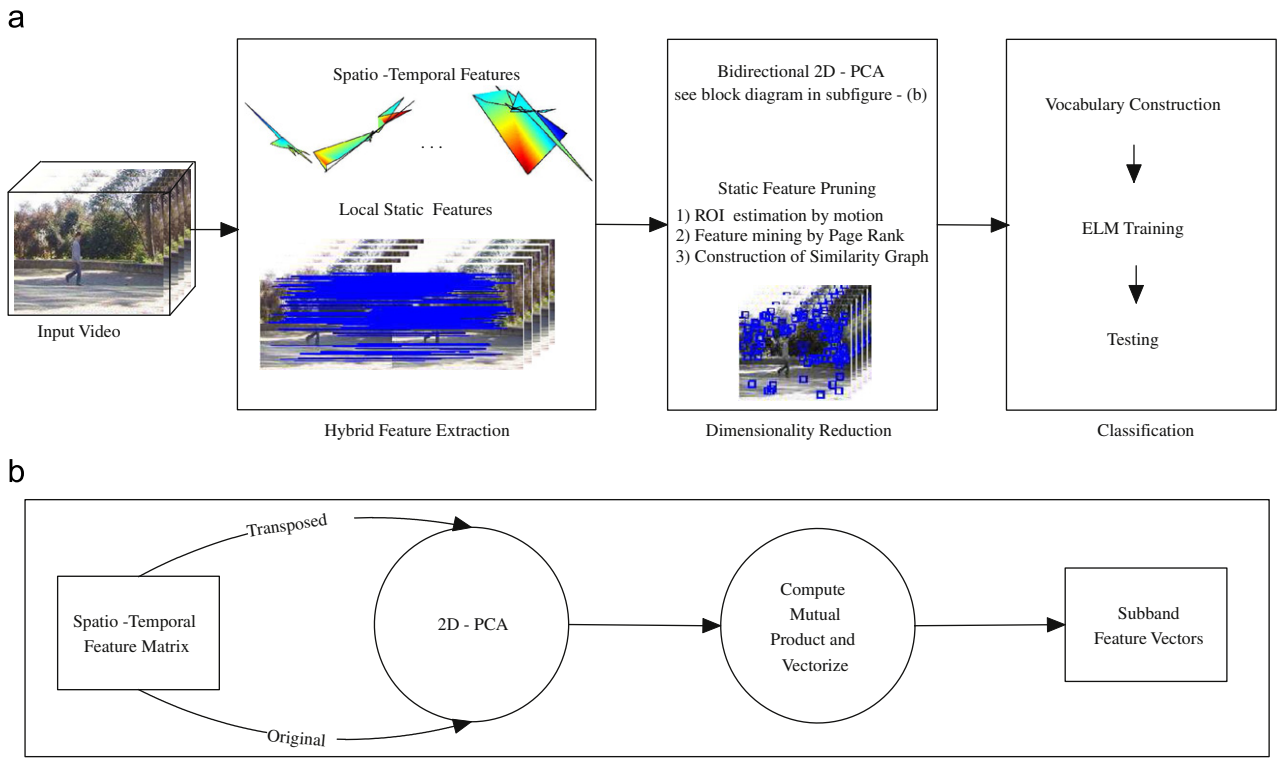

Fig. 4. The block diagram of our proposed algorithm: (a) main steps of the proposed scheme; (b) steps involved in computation of bidirectional 2D-PCA.

can be randomly assigned if the activation function is infinitely differentiable. It is an interesting solution; instead of tuning the entire network parameters such random allocation helps to analytically determine the hidden layer output matrix Υ. For the fixed network parameters, the learning of ELM is simply equal to finding a least-square solution of

$$\|\Upsilon(\hat{w}_1,\ldots,\hat{w}_L,\hat{b_1},\ldots,\hat{b_L})\hat{\beta}-\Gamma\|, \tag{11}$$

$$=\min_{w_i,b_i,\beta}\|\Upsilon(w_1,\ldots,w_L,b_1,\ldots,b_L)\beta-\Gamma\|. \tag{12}$$

For a number of hidden nodes $L \ll N$, Υ is a non-square matrix, the norm least-square solution of above linear system becomes $\hat{\beta}=\Upsilon^*\Gamma$, where Υ^* is the *moore-penrose* generalized inverse of a matrix Υ. It should be noted that above relationship holds for a non-square matrix Υ whereas the solution is straightforward for $N=L$. The smallest training error is achieved using above model since it represents a least-square explanation of a linear system of $\Upsilon\beta=\Gamma$ as

$$\|\Upsilon\hat{\beta}-\Gamma\|=\|\Upsilon\Upsilon^*\Gamma-\Gamma\|, \tag{13}$$

$$=\min_{\beta}\|\Upsilon\beta-\Gamma\|. \tag{14}$$

3. Proposed algorithm

For action recognition, support vector machine (SVM), AdaBoost, k-NN, AdaBoost with multiple instance learning (MIL), temporal boosting, chaotic invariants, and representation of actions in space-time shapes have been proposed in literature. The profound use of multiple types of features is evident from improved detection results [12,13] since they provide complementary information for action recognition. However, trade-off between acquired accuracy and computational time poses a major bottleneck for real-time implementation of these schemes in various applications. In this section, we describe our action recognition framework which utilizes ELM for classification using hybrid data, i.e., dimensionality reduced features set. Such data are obtained from two kinds of features, i.e., *spatio-temporal* features and *local static* features. The proposed framework assigns an action label to an incoming video based upon observed activity. The method is capable of identifying a specific action present in a video utilizing an ELM trained on visual vocabularies constructed using hybrid feature vectors whereas we do not assume any *a priori* information about background, view point, activity and data acquisition constraints.

3.1. Synopsis of proposed framework

The implementation of proposed algorithm starts with the computation of hybrid feature vectors whereas resizing of video frames to square dimension and converting our colored video to the gray space are the only preprocessing operations applied. The 3D DT-CWT is employed to extract coefficients which contain embedded *spatio-temporal* information of volumetric data of different moving objects. To generate distinctive and lower dimension *spatio-temporal* information from videos, bidirectional 2D-PCA is applied on subbands of multiresolution decomposition which results into considerably smaller sized feature vectors. The second class of features, *local static* features, are extracted by applying ASIFT on patches around stable interest points detected using Harris-Laplacian and Hessian Laplacian schemes followed by a pruning strategy [13] to eliminate ASIFT descriptors for immobile parts in the scene which carry no useful information about the sought action. Finally, we construct visual vocabularies using both kinds of features with assigned labels to be input ELM for training purpose. Visual vocabularies are a way to represent features for a classifier that associates query images to the training elements. This approach saves us computational efforts

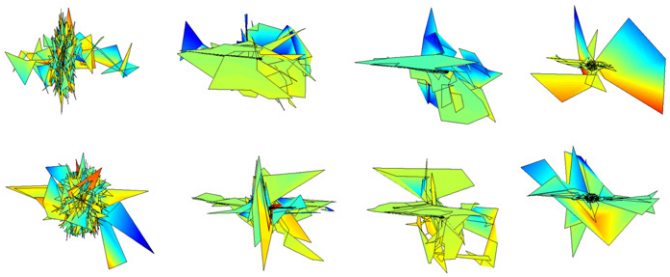

Fig. 5. Some sample spatio-temporal features computed using motion selectivity attribute of 3D DT-CWT. From left to right columns, top view of first directional subband for four actions, namely, bend, run, skip and wave1, respectively.

to relate an incoming image to all training datasets; we try to identify a small number of clusters with excellent discriminative attributes for various classes. A minimized within-cluster and maximized between-clusters scatter is attempted using square-error partitioning, i.e. k-means, which proceeds by iterated assignments of points/features to their closest cluster centers and reevaluation of cluster centers. We do not require background subtraction or object tracking using visual vocabularies and similarity information of features are used to represent relevant video sequences. The block diagram of our proposed algorithm is presented in Fig. 4.

3.2. Spatio-temporal features

For convenience, we use the term of *spatio-temporal* features to refer to subband feature vectors. The *spatio-temporal* feature vectors are extracted from an input video sequence using 3D DT-CWT[1] without any segmentation and stabilization operations. This is an important contribution, the techniques proposed in the past assumed to have knowledge of background or foreground masks or required manual stabilization operation on an incoming video before event recognition [6,26,28]. The motion selectivity attribute of 3D DT-CWT can reliably extract *spatio-temporal* features which are truly discriminative for variations among inter-class and intra-class actions performed by the similar or different actors. Applying 3D DT-CWT on an input 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. Fig. 5 represents extracted features, using first orientational subband decomposition, of four different actions performed by two actors. The top row shows features extracted from action videos of actor Daria whereas bottom row corresponds to actor Shahar. The columns from left to right correspond to four actions i.e. *bend, run, skip* and *wave1*, respectively. It is clearly evident that the extracted *spatio-temporal* features capture important deviations in data occurred because of similar actions performed by different actors under differing dynamics and/or different actions performed by the same actor.

3.3. Bidirectional 2D principal component analysis

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. An image covariance matrix is constructed by directly using the original image matrices. Let X denotes an M-dimensional unitary column vector. To project a $Q/2 \times M/2$ image matrix A on X; a linear transformation $Y=AX$ is used which results in a Q-dimensional projected vector Y. The total scatter of the projected samples is introduced to measure the discriminatory power of a projection vector X. The total scatter can be characterized by the trace of a covariance matrix of the projected feature vectors, i.e., $J(X)=\mathrm{tr}(S_x)$ where $\mathrm{tr}(.)$ represents the trace of a matrix, and S_x denotes the covariance matrix of projected feature vectors. The covariance matrix S_x can be computed as

$$S_x = E[(Y-E(Y))(Y-E(Y))'], \tag{15}$$

$$= E[[(A-E(A))X][(A-E(A))X]'], \tag{16}$$

$$\mathrm{tr}(S_x) = X'[E(A-EA)'(A-EA)]X. \tag{17}$$

Yang et al. [23] showed that extraction of image features using 2D-PCA is computationally efficient and better recognition accuracy is achieved compared with traditional PCA. However, the main limitation of 2D–PCA based recognition is the processing of higher number of coefficients since it works in row directions only. Pang et al. [24] suggested an efficient approach, named binary 2D-PCA, to approximate bases of 2D-PCA using Haar like binary box functions. 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 training image mean need not be necessarily equal to zero. To avoid the curse of dimensionality bidirectional 2D-PCA is employed (see Fig. 4(b) for flow chart of bidirectional 2D-PCA computation). One may come up with two basic questions that why do we need dimensionality reduction and if it is needed then why to use bidirectional 2D-PCA? For first question, we believe that the reduction in dimension of data will enhance training and testing speed of our classifier at later stage, secondly, our extracted feature sets also contain static information, such as background and motionless objects in the scene, given that we do not apply segmentation or stabilization operation on an incoming video. Such stationary information in a feature set causes increased ambiguity and classification complexity which can be minimized by extracting discriminative information using dimension reduction scheme. For later query about why to use bidirectional 2D-PCA instead of any other linear/non-linear diminution scheme, we are mainly interested to retain the correlation amongst adjacent data points which plays an important role in volumetric data of an action for accurate recognition. Fig. 6(a) shows better ability of bidirectional 2D-PCA to represent the spatio-temporal information of various action categories performed by actor Daria. Fig. 6(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. 6(a), the separability of different action classes is noticeable whereas components are merged for the feature vectors obtained using PCA (Fig. 6(b)).

3.4. Local static features

The humans have ability to recognize an action from a collection of instantaneous poses of an object in still images. In such data, only shape and its context information is available whereas the motion interpretation is absent. Shape context, histogram of gradient of local neighborhood, and appearance have profusely been used in problem domains like recognition and classification. For automated action recognition using instantaneous frames, more than one images are required to cope with the unpredictable camera movements.

[1] The filter bank used in 3D DT-CWT is derived from our previous work [37].

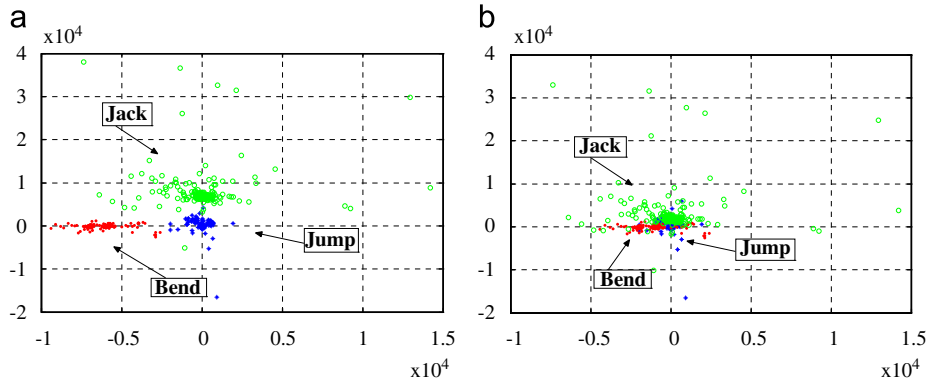

Fig. 6. Distinctive features represented among different videos. Spatio-temporal information captured by (a) bidirectional 2D-PCA and (b) PCA.

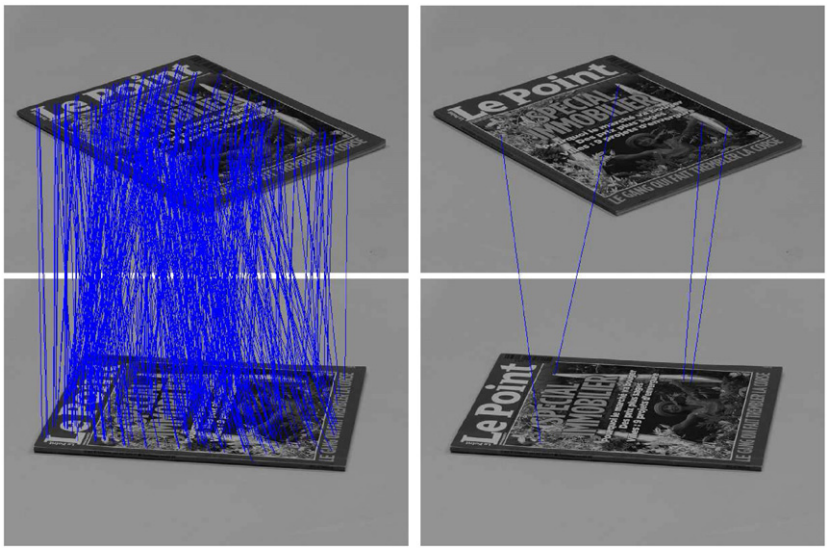

Fig. 7. Matching of image pair using ASIFT and SIFT Methods [14]. *Left*: ASIFT matching, *right*: SIFT matching.

The well known image detectors like SIFT [10], maximally stable extremal region (MSER), level line descriptor (LLD), Hessian-Affine and Harris-Affine are designed to locate interest points in the presence of affine transformations. However, these methods are not fully invariant to scale changes and affine transformations, SIFT performs better than other methods for the images with large scale changes. Affine SIFT (ASIFT) is a recent addition to the family of local image detectors [14] that can reliably identify features which have undergone very large affine distortions (see Fig. 7). ASIFT has improved ability to detect local patches which are distorted by the parameter *transition tilt* upto 36 and higher whereas none of aforementioned methods support this variation above 10. The *local static* features are described using ASIFT descriptor applied on the patches located around interest points identified by Harris-Laplacian and Hessian-Laplacian. The Harris-Laplacian locates corner features while the blob features are identified using Hessian-Laplacian. Both feature types serve as complementary information for each other. Fig. 7 depicts image matching capabilities of ASIFT and SIFT, it clearly validates the claim that ASIFT outperforms SIFT regarding number of correct matches between two images of the same magazine largely distorted by affine transformation. Employing detectors (Harris-Laplacian and Hessian-Laplacian) without segmentation and stabilization operation on video frames has inherent shortcoming to locate interest points which belong to static scene information such as background or stationary objects. We do realize that ASIFT descriptors for patches around such points may not provide any discriminative information for accurate recognition hence such *local static* features are eliminated using pruning strategy proposed by Liu et al. [13]. The pruning operation serves as reduction operation on quantity of *local static* features as presented in 3rd column of Fig. 4(a) where the feature descriptors of patches to be eliminated are bounded in blue squares; it is noticeable that these squares are not associated with moving human body parts hence they do not carry any discriminative information and are justifiable to be removed in further recognition stream.

4. Results and discussion

To test the performance of our proposed method, publicly available datasets, Weizmann human action dataset [6] and KTH dataset [11], are used in our experiments. It is pointed out that the results presented cannot be considered as direct comparison against other recognition schemes because of all kinds of variations of the experimental setups and assumptions about *a priori* knowledge of the video/action being investigated. However,

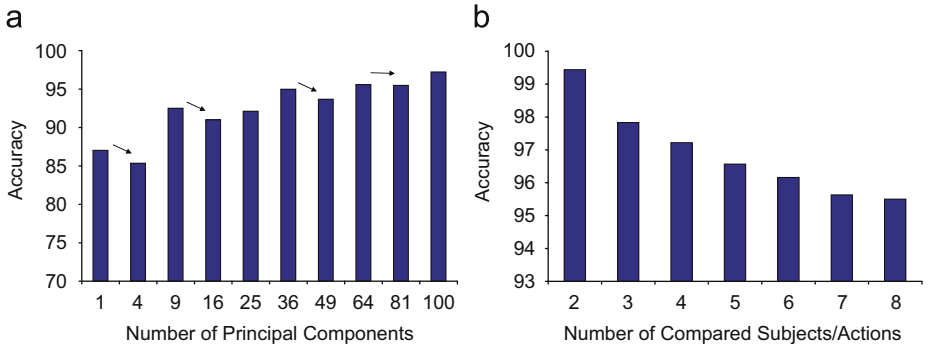

Fig. 8. Accuracy analysis (a) using *spatio-temporal* (with varying size) features only (b) varying number of compared subjects/actions using hybrid features.

the presented results demonstrate that our proposed method is robust and can produce comparable recognition accuracy to other well-documented approaches. Due to lack of an established quality measure protocol, the best reported recognition accuracies from past research are quoted. For simplicity, we present recognition results of only one dataset provided that the similar identification trend is observed for rest of datasets also.

The Weizmman human action dataset [6] contains 83 video sequences showing nine different subjects which perform nine distinct actions at varying speeds. The KTH dataset contains six types of human actions, i.e., *walk*, *jog*, *run*, *box*, *wave* and *clap*. 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. Whereas, we executed all of our experiments in MatLab environment on an Intel Core 2 Duo processor of 1.80 GHz speed and 2 GB RAM.

The important advantages offered by our proposed scheme comprise of no requirement of video alignment and the number of feature vectors is proportional to the number of frames in a video and levels of multiresolution decomposition to extract *spatio-temporal* features. We apply our classification scheme using *spatio-temporal* features only, extracted from Weizmann dataset [6], to demonstrate the need for hybrid features. It is worth pointing that the dimension of individual feature vectors may 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. 8(a), accuracy is not constantly increasing by raising dimensionality of feature vectors; especially classification precision is dwindling or remains constant at arrow locations while the feature vector dimensionality is ever-increasing. Size of feature vectors are mentioned as a square of positive integer value since we are using 2D-PCA based approach in orthogonal directions on 3D DT-CWT coefficients. Fig. 8(a) corroborates our claim that we cannot persistently increase the size of the spatio-temporal features since the accuracy is not promised but computational complexity. It should be noted that *local static* features provide complementary information for accurate recognition because the use of *spatio-temporal* features alone does not guarantee precise identification of an action whereas the selection of optimal size of features is still a mystifying barrier.

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. Our proposed method achieves a varying classification precision for different number and combinations of subjects/videos on Weizmann dataset (see Fig. 8(b)). One subject is randomly selected and its corresponding videos are used as testing set whereas the videos of remaining subjects act as training set; the number of compared subjects, $2 \leq \Lambda \leq 8$, in Fig. 8(b) correspond to the average classification accuracy achieved for Λ number of subject videos being used in testing. The classification accuracy presented in Fig. 8(b) is obtained using hybrid features and it is apparent that the trend line of achieved accuracy is downwards for higher number of compared 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 due to apparently the same activity in between repetitions of an actions is observed in a small number of adjacent frames.

4.1. Why ELM and hybrid feature sets for classification

ELM is a relatively new scheme with potential application to problems requiring real-time classification. It is an appealing study to examine the robustness and performance of proposed framework regarding two important issues: (1) why ELM instead of other classifiers? (2) does the combination of ELM along with hybrid features offer better recognition accuracy? A set of rigorous experiments are performed with varying combinations of classifiers (ELM, AdaBoost and SVM) and various feature sets extracted using benchmark datasets. The spatio-temporal features included in our trials comprise of spatio-temporal (ST), and space-time interest points (STIP) features using three different local space-time descriptors, i.e., HoG, HoF and the combination of both represented as HnF. For all experiments presented in this section, it should be noted that 20 stumps and linear kernel have been used for AdaBoost and SVM classifier, respectively.

For binary classification, three established classifiers (AdaBoost, SVM and ELM) are tested for accuracy and computational complexity using hybrid features extracted from Weizmann dataset. It should be noted that the training and testing features are randomly selected for all iterations of our experiments whereas once selected the similar data is input to all classifier to fairly verify their learning and identification abilities. The selected action videos for each iteration are merely random while the classification setup is also extendable to multi-class problems.

Accuracy of classification is illustrated in Fig. 9, it confirms the improved performance of ELM and AdaBoost over SVM, where ELM and AdaBoost show competitive results with a slightly better performance for ELM. For similar experiment, increasing number of iterations to analyze the statistics, it is seen that the accuracy of ELM is on the average higher than AdaBoost, as shown in

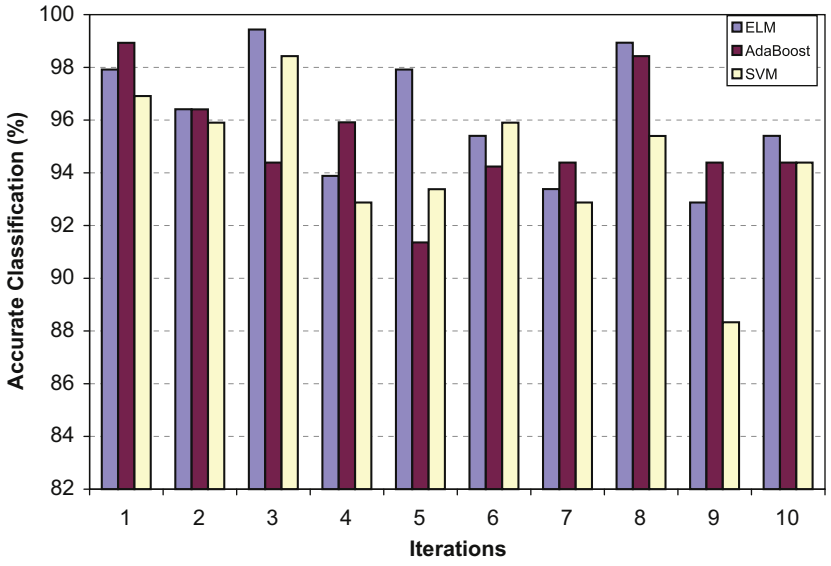

Fig. 9. Accuracy analysis of different classifiers using our hybrid features for Weizmann dataset.

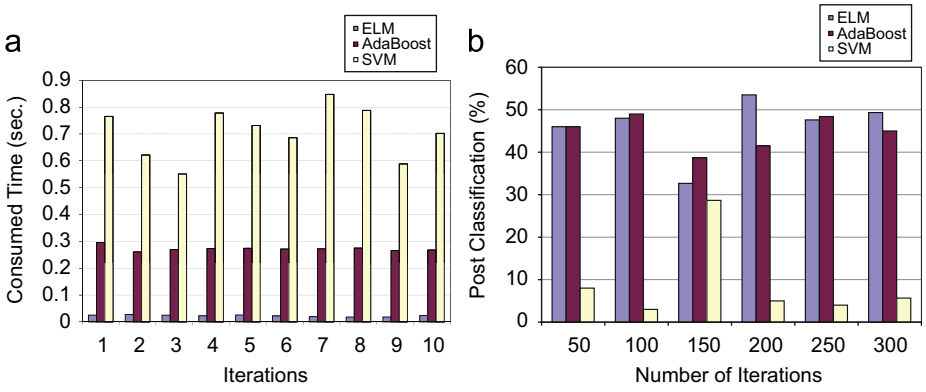

Fig. 10. Performance analysis of different classifiers using Weizmann dataset: (a) computational complexity analysis and (b) best classifications achieved for differing iterations.

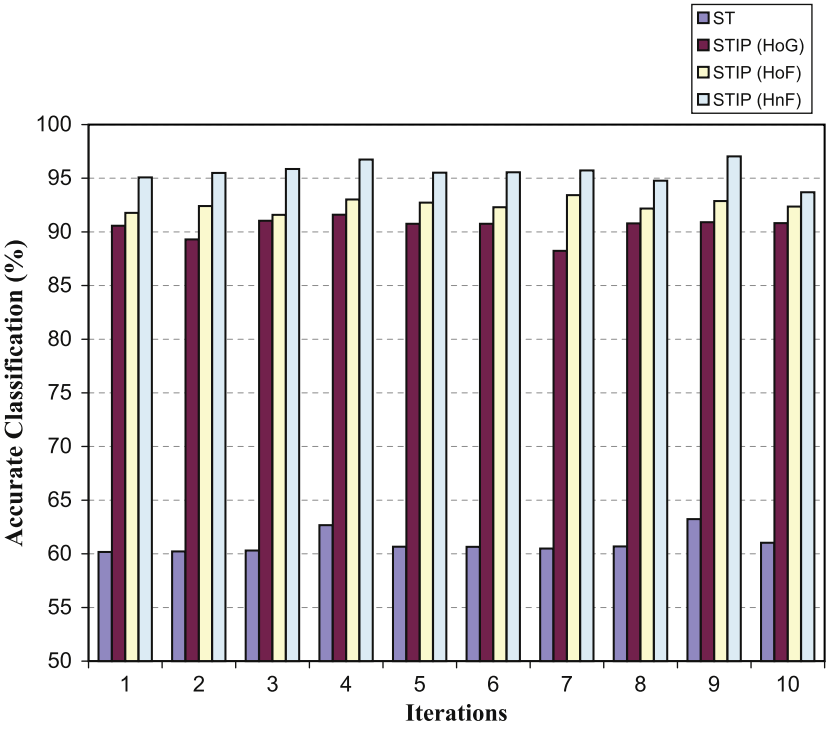

Fig. 11. Performance analysis of ELM using various *spatio-temporal* features for Weizmann dataset.

Fig. 10(b) in terms of collective percentage of best classifications achieved by individual classifiers.

In terms of computational complexity, as shown in Fig. 10(a), SVM is the most time consuming method with a fluctuating behavior. AdaBoost and ELM show a steady (almost) computational time where ELM outperforms the former with a notable time difference. The computational cost becomes an important factor if the number of iterations or size of input data is increased. The lower computational burden and comparable accuracy are the deciding factors for ELM to be used as recognition classifier in our proposed framework.

As a next step, for the similar data sets ELM is employed to four types of spatio-temporal features namely ST, STIP (HoG), STIP(HoF) and STIP (HnF). Fig. 11 depicts the generated accuracies for differing iterations. A persistent behavior and the highest accuracy is achieved using STIP (HnF) with ELM while ST features perform the worst. The classification performance is gradually rising in order of features ST, STIP(HoG), STIP(HoF) and STIP(HnF). For all iterations, the average accuracy for ST features is close to 61% while we are able to achieve an average accurate classification of 95.70% for STIP (HnF) features. The recognition achieved using ELM learned by hybrid features is able to achieve relatively better performance (please refer to discussion below) which substantiates our choice to select ELM and hybrid features together for improved classification.

Table 1
Confusion table of per-video classification for Weizmann dataset [6].

	Bend	Jump	Jack	Side	Walk	Run	Pjump	Wave1	Wave2
Bend	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Jump	0.0	0.99	0.0	0.0	0.0	0.01	0.0	0.0	0.0
Jack	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Side	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Walk	0.0	0.0	0.0	0.01	0.97	0.02	0.0	0.0	0.0
Run	0.0	0.0	0.0	0.0	0.01	0.99	0.0	0.0	0.0
Pjump	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
Wave1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Wave2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Table 2
Confusion table of per-video classification for KTH dataset [11].

	Box	wave	Clap	Jog	Run	Walk
Box	1.0	0.0	0.0	0.0	0.0	0.0
Wave	0.02	0.98	0.0	0.0	0.0	0.0
Clap	0.03	0.01	0.96	0.0	0.0	0.0
Jog	0.0	0.0	0.0	0.88	0.04	0.08
Run	0.0	0.0	0.0	0.06	0.90	0.04
Walk	0.0	0.0	0.0	0.02	0.01	0.97

4.2. Performance analysis of proposed framework

Table 1 presents confusion table, with achieved accuracy of 99.44%, for a random combination of videos used for testing and training purpose, respectively. It can be seen that only three videos, from Weizmann dataset, are partially misclassified. The first confusion in classification is observed for video sequences which are labeled as *running* while actually they belong to *jump* and *walk* actions whereas *run* has also been wrongly recognized as *walk* at some point in recognition process. Apparently, *run–walk* are quite similar actions because they only differ by the speed of a performed action. A *jump* video is misrecognised as *run* which is a hard classification problem since both videos contain an action to pass in front of camera from side view at a faster speed. The last wrongly classified video is *walk* which has been labeled as *side* and *run* since the movements in lower body parts for all three actions are visibly very close.

Furthermore, we test our proposed algorithm using publicly available KTH video sequences [11] for six various actions; from confusion matrix (see Table 2) it is noticeable that classification uncertainty is present among two subgroups of actions, i.e., actions mainly involving hand movements such as *box*, *wave* and *clap* whereas second class of actions consists of legs/feet as major motion elements (*jog*, *run* and *walk*). The action of *jogging* is the hardest classification task because of its similarity with *running* and *walking*; the second most complicated classification chore corresponds to the video sequences of *running* and *clapping*. We are able to achieve the precise classification of 94.83% for KTH datasets, which demonstrates favorable results for proposed action recognition scheme.

Figs. 12 and 13 present performance analysis, for both datasets, i.e., Weizmann and KTH, of various methods for human action recognition; the proposed approach outperforms the previously reported techniques in terms of accuracy whereas our proposed scheme does not require any *a priori* information

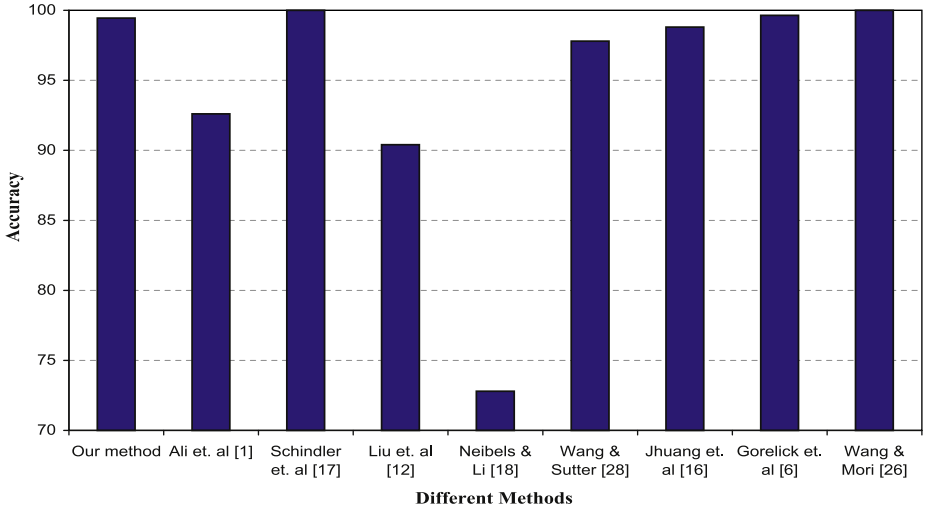

Fig. 12. Performance comparison for various methods using Weizmann datasets [16,18].

such as foreground masks for segmentation. On all video sets, our approach renders improved and/or comparable recognition accuracy against existing schemes. In Fig. 12, the categorization precision of [6,17,26] is slightly better than our scheme, however, the preprocessing steps of these recognition methods require specialized knowledge of the scene being probed. The specialized knowledge may comprise of known background, stabilization of a video sequence that demands only one dynamic object in a frame and video trimming to avoid action repetition that causes misclassification because of similar actions being observed in between action reiterations.

For KTH dataset, our proposed scheme generates recognition accuracy of 94.83% which compares favorably to previous approaches in terms of correctness; please refer to recognition accuracies of [17,26,27,34] in Fig. 13.

4.3. Robustness test

The proposed scheme is tested using Weizmann robustness dataset that consists of various actions performed by the subject(s) inside a room or in an outdoor environment with illumination fluctuations, differing walking styles, multiple moving objects (man walking with dog or minor movements in trees in the background), non-rigid deformations and partial occlusions. The walking action is the most observable movement in a daily life; we test our algorithm on 10 different styles of walking from the same view point. Fig. 14 presents samples video frames from the dataset under-reference where the clutter background, partial obscured human body parts because of skirt, pole, and box in front of feet contribute to complicate the classification task. Another kind of activity videos, termed as *robust view*, are

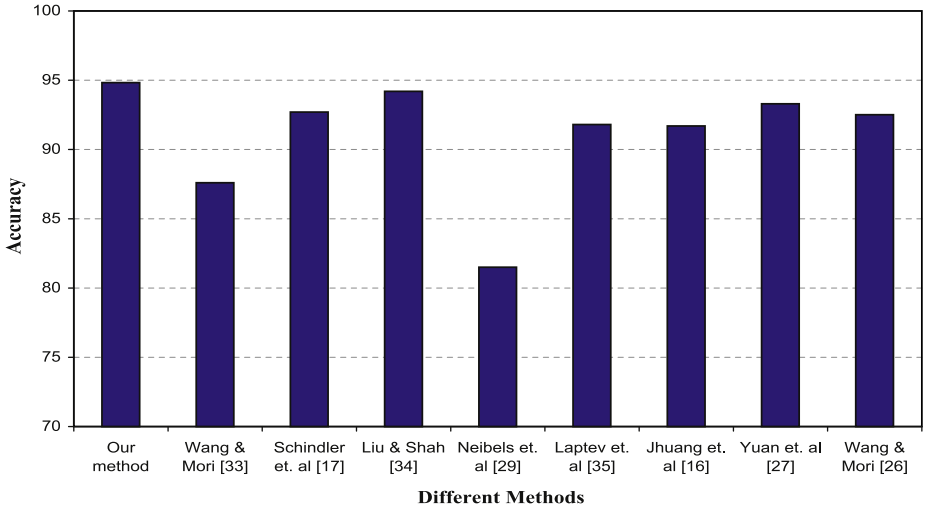

Fig. 13. Performance comparison for various methods using KTH datasets [16,33].

Fig. 14. Sample images from Weizmann robustness datasets. Left to right: *with dog*, *with bag*, *knees up*, *pole*, *in skirt* and *no feet* action videos.

a

Dataset	Varying Parameters	Dataset	Varying Parameters
Robust View	Change in scale and view point	No feet & Pole	Partial occlusion
With bag	Rigid deformation and partial occlusion	Norm walk	Dynamic background
Carry briefcase	Partial occlusion	With dog	Non-rigid deformation
In skirt	Clothes causing extraneous movements	Knees up	Walk style
Moon walk	Walk style with peculiar arms position	Limp	Walk style

b

Robust view	With bag	Carry briefcase	With dog	Knees up	Limp	Moon walk	No feet	Norm Walk	Pole	With skirt

Fig. 15. Robustness evaluation of our proposed method using Weizmann robustness datasets: (a) details of dataset and (b) recognition comparison for different techniques i.e. [6,28] and our method (top to bottom).

also included in our experimental trials which contain normal walk of an actor whose motion is captured from different view points where both scale and view point deformations are involved. Fig. 15(a) represents fundamental details of observed deformations present in datasets used to vigorously analyze the performance of our proposed algorithm.

The proposed scheme is not fully invariant to view changes, however, it exhibits robust behavior in presence of partially occluded objects, scale changes and non-rigid transformations. Fig. 15(b) presents recognition of various methods applied on above mentioned datasets. The black and gray parts of the bars correspond to correct and wrong classification of a video in action recognition, whereas three bars, from top to bottom, show the achieved accuracy employing [6,28] and our proposed method. It is notable that our proposed method generates 100% precise recognition for various deformations except view point changes for which [6] outperforms all other methods; however, the comparison may not be a fair indication of the dazzling performance of [6] which requires *a priori* background information for accurate segmentation.

5. Conclusion

A new human action recognition framework based on multiple types of features is presented. 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 DT-CWT support reduced artifacts and resourceful processing of a video for better quality and well-localized detection of *spatio-temporal* features while *Static local* features are determined using affine SIFT descriptors. Visual vocabularies constructed using both kinds of features 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 to name a few. Both military and industrial applications can potentially benefit from our recognition framework because of its real-time processing and improved precision compared with other well-established schemes.

Acknowledgements

This research has been supported in part by the Canada Research Chair Program, AUTO 21 NCE and the NSERC discovery grant. Authors are thankful to Ivan Laptev and Barbara Caputo for provision of code to extract STIP features. Authors would also like to express their gratitude to editor and anonymous reviewers for valuable suggestions to improve this paper.

References

[1] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: Proceedings of the International Conference on CV, 2007, pp. 1–8.
[2] M.J. Black, Explaining optical flow events with parameterized spatio-temporal models, in: Proceedings of the International Conference on CVPR, 1999, pp. 1326–1332.
[3] M. Brand, N. Oliver, A. Pentland, Coupled HMM for complex action recognition, in: Proceedings of International Conference on CVPR, 1997, pp. 994–999.
[4] T.J. Burns, A non-homogeneous wavelet multiresolution analysis and its application to the analysis of motion, Ph.D. Thesis, Air Force Institute of Technology, 1993.
[5] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing actions at distance, in: Proceedings of the International Conference on CV, 2003.
[6] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space time shapes, IEEE Trans on PAMI (2007) 2247–2253.
[7] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing (2006) 489–501.
[8] H. Jiang, M.S. Drew, Z.N. Li, Successive convex matching for action detection, in: Proceedings of the International Conference on CVPR, 2006, pp. 1646–1653.
[9] N.G. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals, Journal of Applied and Computational Harmonic Analysis 10 (3) (2001) 234–253.
[10] D.G. Lowe, Distinctive image features from scale-invariant key points, International Journal of Computer Vision 60 (2) (2004) 91–110.
[11] I. Laptev, On space-time interest points, International Journal of Computer Vision 64 (2/3) (2005) 107–123.
[12] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: Proceedings of the International Conference on CVPR, 2008, pp. 1–8.
[13] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos "in the wild", in: Proceedings of the International Conference on CVPR, 2009.
[14] J.M. Morel, G. Yu, ASIFT: a new framework for fully affine invariant image comparison, SIAM Journal on Imaging Sciences 2 (2) (2009).
[15] G. Mori, X. Ren, A.A. Efros, J. Malik, Recovering human body configurations: combining segmentation and recognition, in: Proceedings of the International Conference on CVPR, 2004, pp. 326–333.
[16] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: Proceedings of the International Conference on CV, 2007.
[17] K. Schindler, L.V. Gool, Action snippets: how many frames does human action recognition require?, in: Proceedings of the International Conference on CVPR, 2008.
[18] J. Niebels, F.F. Li, A hierarchical model of shape and appearance for human action classification, in: Proceedings of the International Conference on CVPR, 2007, pp. 1–8.
[19] I.W. Selesnick, K.Y. Li, Video denoising using 2D and 3D dual-tree complex wavelet transforms, Wavelet Applications in Signal and Image Processing, SPIE 5207, San Diego, 2003.
[20] I.W. Selesnick, F. Shi, Video denoising using oriented complex wavelet transforms, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), vol. 2, 2004, pp. 949–952.
[21] I.W. Selesnick, R.G. Baraniuk, N.G. Kingsbury, The dual-tree complex wavelet transform—a coherent framework for multiscale signal and image processing, IEEE Signal Processing Magazine 6 (2005) 123–151 software available online: 〈http://taco.poly.edu/WaveletSoftware/〉.
[22] G. Strang, T. Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge, 1996.
[23] J. Yang, D. Zhang, F. Frangi, J.-Y. Yang, Two-dimensional PCA: a new approach to appearance based face representation and recognition, IEEE Transactions on PAMI (1) (2004) 131–137.
[24] Y. Pang, D. Tao, Y. Yuan, X. Li, Binary two-dimensional PCA, IEEE Transactions on Systems, Man and Cybernatics—Part B 38 (4) (2008) 1176–1180.
[25] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: Proceedings of the International Conference on CVPR, 2005, pp. 984–989.
[26] Y. Wang, G. Mori, Max-Margin hidden conditional random fields for human action recognition, in: Proceedings of the International Conference on CVPR, 2009.
[27] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: Proceedings of the International Conference on CVPR, 2009.
[28] L. Wang, D. Sutter, Recognizing human activities from sillhouettes: motion subspace and factorial discriminative graphical model, in: Proceedings of the International Conference on CVPR, 2007.
[29] J. Niebles, H. Wang, F.F. Li, Unsupervised learning of human action categories using spatio-temporal words, in: Proceedings of BMVC, 2006.
[30] P. Dollár, V. Rabaud, G. Cotteral, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of Workshop on Performance Evaluation of Tracking and Surveillance, 2005.
[31] Y. Yuan, Y. Pang, J. Pang, X. Li, Scene segmentation based on IPCA for visual surveillance, Neurocomputing (2009) 2450–2454.
[32] C. Schüldt, I. Lepatev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the International Conference on PR, 2004.
[33] Y. Wang, G. Mori, Learning a descriminative hidden part model for human action recognition, in: NIPS, 2008.
[34] J. Liu, M. Shah, Learning human actions via information maximization, in: Proceedings of the International Conference on CVPR, 2008.
[35] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the International Conference on CVPR, 2008.
[36] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, H.-J. Zhang, Human gait recognition with matrix representation, IEEE Transactions on Circuits and Systems for Video Technology 16 (7) (2006) 896–903.
[37] R. Yu, A. Baradarani, Sampled-data design of FIR dual filter banks for dual-tree complex wavelet transforms, IEEE Transactions on Signal Processing 56 (7) (2008) 3369–3375.

Rashid Minhas is a Ph.D. Candidate at computer vision and sensing systems laboratory, department of electrical engineering, University of Windsor Canada. He completed B.Sc. Computer Science from BZU Multan Pakistan and MS Mechatronics from GIST Korea. His research interests include object and action recognition, image registration and fusion using machine learning and statistical techniques.

Aryaz Baradarani is currently a Ph.D. candidate at the University of Windsor. He is the recipient of University of Windsor IGE and IGS Scholarships, 3M Company Bursary Award in 2009, and President's Excellence Scholarship. His current interests are mainly in theoretical signal processing and its applications in image processing.

Sepideh Seifzadeh is currently a M.S. student in Computer Science at the University of Windsor. She is the recipient of the 2009–2010 University of Windsor IGS Scholarship, and 3M Company Bursary award in 2009. Her current interests are in robotics and pattern recognition.

Q.M. Jonathan Wu (M'92, SM'09) received his Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990. From 1995, he worked at the National Research Council of Canada (NRC) for 10 years where he became a senior research officer and group leader. He is currently a Professor in the Department of Electrical and Computer Engineering at the University of Windsor, Canada. Dr. Wu holds the Tier 1 Canada Research Chair (CRC) in Automotive Sensors and Sensing Systems. He has published more than 150 peer-reviewed papers in areas of computer vision, image processing, intelligent systems, robotics, micro-sensors and actuators, and integrated micro-systems. His current research interests include 3D computer vision, active video object tracking and extraction, interactive multimedia, sensor analysis and fusion, and visual sensor networks.

Dr. Wu is an Associate Editor for IEEE Transaction on Systems, Man, and Cybernetics (part A). Dr. Wu has served on the Technical Program Committees and International Advisory Committees for many prestigious conferences.