2010

An Analytic Training Approach for Recognition in Still Images and Videos

Rashid Minhas
University of Windsor

AN ANALYTIC TRAINING APPROACH FOR RECOGNITION IN STILL IMAGES AND VIDEOS

by

Rashid Minhas

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Electrical and Computer Engineering
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy at the
University of Windsor

Windsor, Ontario, Canada

2010

An Analytic Training Approach for Recognition in Still Images and Videos

by

Rashid Minhas

APPROVED BY:

Dr. John S. Zelek
University of Waterloo, Canada

Dr. Robin Gras
School of Computer Science

Dr. Maher Sid-Ahmed
Department of Electrical and Computer Engineering

Dr. Chunhong Chen
Department of Electrical and Computer Engineering

Dr. Q.M. Jonathan Wu
Department of Electrical and Computer Engineering

Dr. Hanna Maoh, Chair of Defense
Faculty of Graduate Studies

3rd May, 2010

Co-Authorship Declaration

I hereby declare that this dissertation incorporates the material that is the result of a joint research, as follows:

This dissertation incorporates the outcome of a joint research undertaken in collaboration with Dr. Aryaz Baradarani, Sepideh Seifzadeh and Abdul Adeel Mohammed under the supervision of Dr. Q.M. Jonathan Wu. The collaboration is covered in Chapters 2-3 of the dissertation. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contributions of co-authors were primarily through the provision of proof reading, software design and reviewing the research papers regarding the technical content.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my dissertation, and have obtained a written permission from each of the co-authors to include the above materials in my dissertation.

I certify that, with the above qualification, this dissertation, and the research to which it refers, is the product of my own work.

Declaration of Previous Publications

This dissertation includes following four papers that have been previously published in peer reviewed journals and conferences.

Chapter	Publication title/full citation	Status
Chapter 2	R. Minhas, A.A. Mohammed, Q.M. J. Wu. "A Fast Recognition Framework Based on Extreme Learning Machine Using Hybrid Object Information"; Neurocomputing 73 (2010) 1831–1839	Published
Chapter 3	R. Minhas, A. Baradarni, S. Seifzadeh, Q.M. J. Wu. "Human Action Recognition Using Non-separable Oriented 3D Dual Tree Complex Wavelets"; The Ninth Asian Conference on Computer Vision, China, (2009)	Published
Chapter 3	R. Minhas, A. Baradarni, S. Seifzadeh, Q.M. J. Wu. "Human Action Recognition Using Extreme Learning Machine Based on Visual Vocabularies"; Neurocomputing 73 (2010) 1906–1917	Published
Chapter 3	R. Minhas, A. Baradarni, S. Seifzadeh, Q.M. J. Wu. "Human Action Recognition Using Extreme Learning Machine via Multiple Types of Features"; SPIE Conference on Defense, Security and Sensing (2010)	Published

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my dissertation. I certify that the above material describes the work completed during my registration as a graduate student at the University of Windsor.

I declare that, to the best of my knowledge, my dissertation neither infringes upon anyone's copyright nor violates any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my dissertation.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

This dissertation proposes a general framework to efficiently identify the objects of interest (OI) in still images and its application can be further extended to human action recognition in videos. The frameworks utilized in this research to process still images and videos are similar in architecture except they have different content representations. Initially, global level analysis is employed to extract distinctive feature sets from an input data. For the global analysis of data the bidirectional two dimensional principal component analysis (2D-PCA) is employed to preserve correlation amongst neighborhood pixels. Furthermore, to cope with the inherent limitations within the holistic approach local information is introduced into the framework. The local information of OI is identified utilizing FERNS and affine SIFT (ASIFT) approaches for spatial and temporal datasets, respectively. For supportive local information, the feature detection is followed by an effective pruning strategy to divide these features into *inliers* and *outliers*. A cluster of *inliers* represents local features which exhibit stable behavior and geometric consistency.

Incremental learning is a significant but often overlooked problem in action recognition. The final part of this dissertation proposes a new action recognition algorithm based on sequential learning and adaptive representation of the human body using Pyramid of Histogram of Oriented Gradients (PHOG) features. The changing shape and appearance of human body parts is tracked based on the weak appearance constancy assumption. The constantly changing shape of an OI is maximally covered by the small blocks to approximate the body contour of a segmented foreground object.

In addition, the analytically determined learning phase guarantees lower computational burden for classification. The utilization of a minimum number of video frames in a *causal* way to recognize an action is also explored in this dissertation. The use of PHOG features adaptively extracted from individual frames allows the recognition of an incoming action video using a small group of frames which eliminates the need of large look-ahead.

to my

MOTHER and FATHER

with love

Acknowledgements

First of all, thanks to almighty Allah for his countless blessings and bounties on me. I am writing this part with particular inspiration to show my sincere gratitude for everyone who contributed to keep my spirits high throughout the course of my stay at UWindsor. I would like to express my deep-felt appreciation to my advisor, Dr. Jonathan Wu for his advice, encouragement, enduring patience and constant support. He has always been open to suggestions, and offer me plenty of time and space to think and *rethink* about the ideas and research sometimes I felt lost into. His never ceasing belief, and patience during initial years of PhD made me realize the fact that being advisor requires more than discussing about new ideas *only*.

I also wish to thank the other members of my committee, Dr. Maher Sid-Ahmed, Dr. Chunhong Chen, Dr. Robin Gras and Dr. John Zelek. Their suggestions, comments and additional guidance were invaluable to the completion of this work. In arduous working hours, the support and important discussions with Abdul Adeel Mohammed, Mohammed Iqbal, Aryaz Baradarni, Abdul Baqi and Stephanie Warren made me smile and do the laborious things again.

Additionally, I want to thank staff and my group fellows at department of ECE for all their hard work and dedication, providing the means to achieve my long-dreamt objective. And finally, I must appreciate my family for putting up with me during the development of this work with continuing, loving support and no complains. I do not have words to express all my feelings here, only that I love you.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

OI	Object of Interest
ASIFT	Affine Scale Invariant Feature Ttransform
2D-PCA	Two-Dimensional Principal Component Analysis
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
CWT	Complex Wavelet Transform
DT-CWT	Dual-Tree Complex Wavelet Transform
ELM	Extreme Learning Machine
HoG	Histogram of Oriented Gradients
HoF	Histogram of Optical Flow
PHOG	Pyramid of Histogram of Oriented Gradients
BVW	Bag-of-Visual-Words
RANSAC	RANdom SAmple Consensus
MSER	Maximally Stable Extremal Regions
GLOH	Gradient Location and Oriented Histogram
STIP	Space-Time Interest Points
ST	Spatio-Temporal
SVM	Support Vector Machines
HMM	Hidden Markov Models
GTDA	General Tensor Discriminant Analysis

FNN	Feedforward Neural Network
BP	Back Propagatoin
SGBP	Stochastic Gradient descent BP
MIL	Multiple Instance Learning
LLD	Level Line Descriptor
PR	Page Rank
ROI	Region of Interest

List of Symbols

$\mathbf{x_i}$	Input sample i
$\gamma_\mathbf{i}$	Labeled output i
$\mathbf{g}(.)$	Nonlinear activation function
$\mathbf{w_i}$	Weight vector between input and hidden layer
$\beta_\mathbf{i}$	Weight vector between hidden and output layer
$\boldsymbol{\Upsilon}$	Hidden Layer output matrix
$\mathbf{L}$	Number of hidden neurons
$\mathbf{b_i}$	Threshold of i^{th} hidden node
$\boldsymbol{\Gamma}$	Output matrix
$\mathbf{tr}(\mathbf{S_x})$	Trace of the covariance matrix S_x
$\mathbf{J}(\mathbf{x})$	Total scatter criterion
$\mathbf{F_k}$	k^{th} Fern
$\phi_\mathbf{f}$	Scaling function
ϵ	Error to be minimized
$\mathbf{sign}()$	Signum function
$\widetilde{\mathbf{\Psi}}()$	Fourier transform of wavelet functions
ζ	Similarity matrix
$\mathbf{W}$	Adjacency matrix
$\mathbf{D}$	Degree matrix
$\mathbf{G}(\mathbf{V}, \beth, \mathbf{W})$	Similarity graph with V vertices, $\beth$ edge set

$\mathbb{L}$	Graph Laplacian
$\mathbf{F_t}$	Frame at time instance t
$\alpha, \mathbf{b}$	Scaling factor and indicator vector, respectively
$\mathbf{Pr_t}$	Page rank for t^{th} frame
$\mathbf{\Lambda}$	No. of subjects/video
χ	Track window
ς	Rectangular blocks to approximate contour of an OI
$\mho$	Intensity histogram

Chapter 1

Introduction

In general, a scene is a collection of objects which may or may not be interacting with each other. The ability of human beings to recognize objects and scenes is a much researched topic across various scientific fields. These researchers share a common goal to discover the secrets behind successfulness of human vision system. The computer vision community is faced with the challenge of devising novel, robust and efficient algorithms to learn models which are helpful in categorizing huge amount of visual data. The main objective for this dissertation is to develop new techniques and tools to recognize objects of interest (OI) from input data of varying dimensionality, i.e., images and videos. Throughout this dissertation, the terms *object recognition* and *action recognition* refer to the identification of OI in input images and videos, respectively. Ideally, an object recognition scheme should reliably formulate the classification task of locating an OI. The classification formulation allows the utilization of various feature extraction and machine learning techniques in order to learn optimal model from a training dataset. Object recognition has varied applications including robot guidance and automation, path planning, and biometrics. The localized recognition of the target object, robot guidance and path planning are important for the tasks which are too dangerous for humans to carry out and require high precision such as mining, battlefields, and space applications. Figure 1.1 shows some scene

Figure 1.1: Sample images of Caltech dataset representing different variations of each category (shown row wise).

images from the publicly available Caltech dataset with different objects at various scales, lighting conditions and viewpoints.

The identification of actions in videos is far more complicated in comparison to still images due to changing backgrounds, poses, lighting conditions, action dynamics, occlusions and camera movements. The acquisition of non-static but repeatedly identifiable features is of key importance for reliable classification. The typical application of action recognition may include, but is not limited to, surveillance, security, sports events, military applications and the onset prediction of critical events and abnormal behavior at public places. Figures 1.2,1.3 demonstrate some examples from frequently used action datasets Weizmman [6] and KTH [11] acquired with static background and camera conditions. Such controlled environment videos have been used as benchmark input sets to validate the performance of various frameworks. However, researchers

Figure 1.2: First row: Randomly selected actions. *Bending, Running, Jack, Jump, Side,* and *Skip* (left to right). Second row: Segmented frames using [48].

have recently started using videos of unconstrained environment to mimic real life situations. The fundamental concept of this dissertation is to address the challenging but an important problem of recognition of natural human actions in diverse and realistic video settings. Unconstrained videos contain significant camera movements, occlusions, cluttered backgrounds and multiple movements along with a large degree of affine deformations (Figure 1.4). The need for unconstrained environment videos arose because of the limited action classes recorded with simplified scene settings. However, research investigating natural videos with human actions subjected to indi-

Figure 1.3: Sample frames for six different actions (left to right: *boxing, clapping, waving, running, jogging, and walking*)from KTH dataset [11].

vidual variations of people in expression, posture, motion and clothing is limited due to the unavailability of realistic and annotated video datasets [35].

1.1 Basics of Recognition Framework

Action recognition from videos and object recognition in images share common problems like cluttered background, occlusions and varying lighting conditions and, thus, share common strategies to deal with significant intra-class variations. A traditional recognition framework consists of a source, transducer/sensor, a feature detection stage and a classifier to learn a model in a supervised or unsupervised fashion (as shown in Figure 1.5). The *source* produces patterns which might be controlled, random physical phenomenon or acts of nature. The *sensor* generates information in the form of scalar values of vectors about the patterns emanating from the source. Ideally, the sensor should be able to extract components to fully represent source pat-

Figure 1.4: Realistic actions from three classes of human actions: *getting out of car, answering a phone and kissing* (modified version of [35]).

terns to make a decision. The selection of a sensor is based on the required accuracy, available resources and physical limitations of the scene being probed. The third step in recognition is the extraction of features by transforming sensor measurements into the feature space. Pre-processing, interest points computation and dimensionality reduction are commonly used techniques to achieve simple yet distinct features. The last and final block of Figure 1.5 is the heart of recognition framework which provides mapping between feature space and decision value a particular feature belongs to. In general, feature extraction and classifiers play a pivotal role in robust performance of a proposed framework. It should be noted that recently published literature is based on the use of non-linear classifiers or the bank of linear classifier to deal with multiclassification problems.

1.2 Motivation

In this section, a brief review of motivation and the proposed ideas of various recognition approaches are presented. An image or video can be considered as a collection of different blocks such as persons, objects, animals and structures etc. Sometimes this collection shows redundant information which can be removed by representing the original data into new space by using tools like discrete cosine transform (DCT), or principal component analysis (PCA) [49],[50]. Such transformation reduces the dimensionality of the input data, however, correlation amongst neighboring pixels is lost. This leads to an important but mostly overlooked problem: *how to efficiently retain correlation information amongst neighboring pixels* lost during customary dimensionality reduction operations such as PCA and Kernel PCA. In recent research

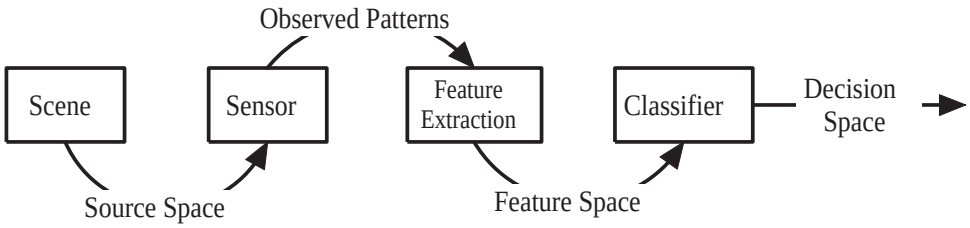


Figure 1.5: Basic setup of a recognition framework.

reports, the bag of features based approaches have been used to address problems related to visual recognition. The bag of features corresponds to clustering the histograms of identified interest points or vectors. The simple yet improved occluded object recognition generated by bag of features prompted its use in text categorization, action recognition and for biometric applications [51],[87]. In literature, the term bag-of-visual-words (BVW) has been used to relate the same concept for features of images and video. With the emergence of new invariant feature detectors, the BVW schemes offer advantages like affine invariance and lower memory requirements since the feature labels are stored instead of p-dimensional vectors. In addition, the BVW offers tolerance to both inter-class and intra-class deformations and lower

computational burden in matching because of the quantization of alike features into similar clusters. The negative side of the BVW approach is the determination of an optimal vocabulary size. In general, larger vocabulary sizes generate higher accuracy but at the same time they are more prone to be sensitive towards noise. The BVW approaches have inherent limitations of ignoring the mutual spatial and/or geometric information of identified features which leads to inefficient representation of the global structure of an object or action being sought. Hybrid features are being increasingly used due to the complementary information offered which improves recognition accuracy [86],[95],[12]. The efficient detection and representation of *global* as well as *local* features guarantees classification to be carried out in real-time since the major computational burden during the testing phase is posed by feature detectors which can be minimized by using adaptive region of interest selection.

The final step of recognition consists of efficient classifiers that can efficiently handle linear as well as non-linear mapping amongst features and their classes. In the past, k-NN, AdaBoost, SVM, HMM, neural networks and probabilistic models have been used for classification and correspondence problems [86],[87],[12],[13],[26]. Traditionally, the performance of a classifier is mainly dependent upon the training phase to learn a model. In recognition research, two potential problems involving the training of a classifier have not gained much attention despite their direct effect on real-time implementation of the existing frameworks. First of all, much less research efforts have been devoted to minimize the training time which poses a major bottleneck in classification. Secondly, can the training be performed in an sequential way while maintaining lower computational complexity?

After thorough review of literature it is observed that action recognition research is mainly concentrated on batch mode processing. In batch mode processing, it is assumed that the entire video is stored in advance before being labeled for a specific action or relatively large look-ahead is available to identify an action present in each frame. There is a need for research which proposes a solution to *label an action present*

in video frames based on minimum number of frames. This solution will not rely on assumptions about background, foreground masks, position of an object of interest and any unrealistic pre-processing like video-level normalization and stabilization. Additionally, it should not assume the availability of future frames for identification of an event in an incoming video.

1.3 Contribution

This section briefly describes the approaches developed to deal with the limitations of the existing schemes mentioned above.

- A recognition framework trained analytically leading to learning speed approximately thousands times faster than traditional learning paradigms.
- A dimensionality reduction technique to better preserve correlation information amongst neighboring pixels or coefficients of feature space.
- An improved framework utilizing parallel classifiers, i.e., extreme learning machines (ELMs) to simultaneously process hybrid feature sets determined using unique *global* and *local* information of an input data.
- In action recognition, an automatic removal of instable and useless *local* features that belong to background or static areas of the scene using principal eigenvectors of graph Laplacian is employed.
- A new class of *spatio-temporal* features selected using the motion selectivity attribute of 3D dual-tree complext wavelet transform (3D DT-CWT) with better ability to capture the dynamics of inter-class and intra-class variations.
- For larger deformations, computation of *local* features in images and videos using FERNS and affine SIFT (ASIFT), respectively, to find correspondence between local patches.

- A real-time action recognition without look-ahead information to segment or track an object to label a video.
- An incremental training approach to adaptively train the classifier for unseen instances of the action being probed.

1.3.1 Object Recognition using Analytically Trained Classifier

To improve the previously proposed recognition framework based on partial object information [95], a new supervised recognition scheme is presented which employs hybrid features, i.e. *global* and *local*, for accurate classification of an object at a considerably higher speed [52]. The global and local object information is extracted using bidirectional two-dimensional PCA (2D-PCA) and Ferns based conditional probabilities, respectively, while parallel extreme learning machines (ELMs) are employed to classify individual feature types. Finally, a fusion process is initiated to integrate classification estimates of all ELMs based on a normalized weighted sum strategy. The first contribution of this research is unique image representation using bidirectional 2D-PCA and the Ferns style approach to represent *global* and *local* feature sets, respectively. The second contribution is the application of ELM, a single hidden layer feedforward neural network which transforms 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 transformation supports reliable recognition with minimum error and at learning speed thousands of times faster than the traditional neural networks. The simpler structure of the classifier enables the categorization task to be finished in fraction of seconds which is significantly faster compared to other modern algorithms [74, 87]. The superior performance of proposed method is observed through comparable accuracy against state-of-the-art approaches.

1.3.2 Action Recognition using 3D Dual-Tree Complex Wavelet Transform

The proposed action recognition technique introduces an efficient way for simultaneous processing of multiple video frames to extract spatio-temporal features for finer activity detection and localization [54]. These features are obtained through the use of motion-selectivity attributes of 3D dual-tree complex wavelet transform (3D DT-CWT) and are used to train a classifier for categorization of an incoming video. The proposed learning model offers three core advantages: 1) the proposed learning framework is trained significantly faster than traditional supervised approaches, 2) The use of the 3D transform allows simultaneous processing of volumetric video data instead of frame by frame analysis, 3) richer representation of human actions because of the directionality and shift-invariant property of DT-CWT. Isolating motion into several subbands in different directions, and celebrated properties of non-separable 3D DT-CWT reduces artifact generated by separable 2D transforms. No assumption of scene background, location, objects of interest, or point of view information is required for activity learning. Bidirectional 2D-PCA is employed for feature extraction that has enhanced capabilities to preserve structure and correlation amongst neighborhood coefficients of a video frame. The spatio-temporal features extracted using 3D DT-CWT provide improved representation of governing dynamics involved to perform a particular action. For classification, ELM is utilized because of its expeditious training capabilities and generalized performance. The research results of this dissertation compare favorably to recently published results in literature.

1.3.3 Visual Vocabularies for Action Recognition

Most of existing frameworks use only the global information of the input video after segmenting moving objects for action recognition. Despite favorable results the computational load to segment a foreground object and extract spatio-temporal fea-

tures is relatively high. The 3D DT-CWT based scheme for action recognition is further refined through incorporation of local information of a moving person to improve accuracy [55],[105]. The new technique introduces a human actions recognition framework based on multiple types of features without the need of a segmentation phase. Taking the advantage of the motion-selectivity property of the 3D DT-CWT and the affine SIFT local image detector, spatio-temporal and local static features are extracted. The framework does not assume any unrealistic pre-processing (like stabilization and/or foreground masks to track a movement) for an incoming video. Intuitively, all of the local descriptors do not carry discriminative information related to an action being detected since features may also be detected from static objects or background of a scene. All such features with low information are required to be eliminated in further recognition processes. Pair-wise constraint of features are applied to prune such features which are instable and not detected throughout the video. Such constraint is helpful to discover the discriminative foreground features where the matching is performed only on a pair of video frames which are not adjacent. Finally, visual vocabularies of both kinds of features are generated to be used for training of an ELM. The proposed technique is significantly faster than traditional methods due to volumetric processing of input video, and offers a rich representation of human actions in terms of reduction in artifacts. Experimental examples are provided in the sections below to illustrate the effectiveness of the proposed approach. Both military and industrial applications can potentially benefit from the proposed recognition framework due to its real-time processing and improved precision as compared to other well-established schemes.

1.3.4 Action Recognition through Recursive Training

Previous section described action recognition schemes using multiple types of features extracted from entire video which presents a batch mode solution of the problem in-hand. To cope with practical situations the *common* assumption of availability of

future frames in action recognition has to be ignored. A limited number of existing action recognition techniques are based on features extracted from past video frames without specialized pre-processing [17]. A unique framework is proposed for real time action recognition based on moving object tracking through online updating of appearance and shape. The changing shape of a moving object is modeled through a small number of rectangular blocks. The pyramid histograms of oriented gradients (PHOG) [57] are computed of these blocks to represent global and local parts of the human body for changing locations with the course of the video. However, the rectangles bounding local parts of the human body may overlap leading to redundant but related information. Both kinds of features are learned using the incremental training scheme by providing sets of training data to ELM while the number and size of training sets need not be necessarily pre-defined. For the incremental learning scheme of ELM, except for the number of hidden nodes no other network parameters need to be manually selected. The proposed scheme offers real-time implementation of action recognition without any segmentation or stabilizing operations on an input dataset. Furthermore, the features are adaptively extracted from the scene regions best representing the changing shape and appearance of a moving object.

1.4 Literature Survey

It is noticeable that recognition techniques for objects and actions carry many common characteristics in modeling and classification. For the last three decades, research in visual recognition has focused on information and feature extraction from 2D images. Just like other fields, initial investigation in object recognition started with assumptions to keep things simple such as uniform background, however, the growing requirements and applications of this area led to the development of methods which are far more complicated and natural in essence. A review of recent work on recognition in still images and videos is presented in the next two sections.

1.4.1 Related Work in Object Recognition

Visual representation is of fundamental importance in recognition frameworks. The projection of a 3D world onto 2D planes adds complications in recognition and the situation is more aggravated due to lack of intelligence in machines today. Li et al. [73] present an excellent introduction to the challenges and recent advancements in learning and recognition of single and multiple classes of objects. Object recognition methods can be mainly divided into two main categories 1) geometry based methods 2) appearance based methods. Existing techniques from both categories are briefly reviewed below.

Geometry based Object Recognition

Due to limited computation power in the past, recognition schemes assumed problems in controlled environments with stable illumination and background. Early object recognition methods focused on identification of an object using range data [58],[59] which provides better depth information of the scene and can ultimately support reliable detection of contours and regions. Image intensity values have been pointed out as the prime source of information due to passive nature of visual sensing devices and advancement in vision research. However, automated extraction of features is still an open problem for the scenes with cluttered background and various deformations.

Simply, a recognition scheme can be defined as a strategy to search an object in 2D projection i.e. an image. One category of object recognition operates on the *hypothesis-proof* principle, first is the hypothesis stage where a correspondence is found between a post of the 3D object model and image features and later the model is projected onto the image and all the evidence is used to verify the judgement [60]. Fischler proposed an efficient algorithm, RANdom SAmple Consensus (RANSAC), to reduce the number of hypotheses [106] whereas trees have also been used to integrate geometric constraints among primitives and explore all possible correspondences to speed up the search [107]. The basis idea of RANSAC is to compute the aligning transformation using a minimal number of randomly sampled correspondences. The

degree of its consensus with other correspondences is used to measure the accuracy of one transformation. Another subcategory of object recognition based on geometrical information, a set of features are extracted and matched against the hash table. This type of indexing has successfully been implemented for document indexing.

All geometry based methods are helpful to recognize an object by matching images to 3D objects which is feasible due to geometric constraints subject to the availability of the discriminative shape of an object in real images. However, such information is not always available which leads to the utilization of the appearance based methods for recognition.

Appearance based Object Recognition

Most real images are composed of multiple objects of various categories with cluttered backgrounds which complicates successful implementation of geometric constraints. In such situations, the modeling of low-level features is easier and more reliable. Often, the knowledge of object profile is insufficient to identify them in images whereas the appearance of the same object may provide more meaningful and identifiable information. The appearance based methods can basically be divided into two main classes 1) pixel-level methods and 2) patch-level method.

a) Pixel-level Methods

This category is based on histograms or statistics of raw features such as output of a filter operation at single pixel. The histogram intersection can also be used for partial matching in situations like occlusions. A simpler way to represent statistics of raw features is the use of histograms, spatial correlation of colors, histogram of oriented gradients (HOG) and pyramid of HoG (PHOG) [61],[109],[57]. It should be noted that raw features may include very simple features such as gradient, pixel values, textures and colors. For efficient processing, coarse level feature can also be used as a starting point to detect in high resolution images with lower computational complexity.

Another powerful set of appearance features is the Haar-like features for object

recognition [77] which is named for its intuitive similarity with Haar wavelets. Texture features have also been used in recognition schemes where the computation of texture features is performed using Wavelet transform, Gabor filters, steerable features and gray-level co-occurrence method [62], [49], [110].

b) Patch-level Methods

These methods are computationally efficient as compared to pixel-level methods since only a small number of patches need to be processed instead of a higher number of pixel values. Secondly, patch-level detectors have been effectively used for recognition because of their repeatable and robust performance in presence of large deformations. Recent advancements in object recognition have produced amazingly accurate classifications based on partial object information represented through patch level descriptors. The fundamental concept is to represent patches selected randomly or around interest points by a vector, such as the Scale Invariant Feature Transform (SIFT) [10], which can be directly compared against other vectors or their clusters [86],[95].

A variety of feature detectors are available ranging from corners, and edges [111]-[115]. The corners are the most commonly used feature in object recognition which can be detected using Harris corner [116] and other techniques [56],[111],[112]. Schemid proposed local gray-value invariants for image retrieval. However, the Harris corner detector is not invariant to scale changes. To cope with such image deformations, feature detectors can be applied on various image scales. Edges have profoundly been used in recognizing objects of different classes [74],[97],[104]. These techniques not only devise a shape model for contours as more discriminative, but also apply contour matching algorithm for recognition. To capture both appearance and contour information, contour networks and the extraction of patches from sampled points of contours have been proposed [74],[117]. Most detectors apply differential or intensity extrema strategy; a sequence of nested contiguous region is obtained using various thresholds on an image and later Maximally Stable Extremal Region (MSER) is applied to find the regions with approximately stationary area. Different variants of

basic feature detectors are invariant to translation, rotation and scaling. For robust performance, research efforts are directed towards the development of detectors which are also invariant to perspective distortions such as Harris-affine, Hessian-affine and affine SFIT (ASIFT) to name a few [14],[113].

The patches surrounding detected interest points can be further represented in a discriminative way which is also invariant to different changes. A simple way to find a descriptor which is invariant to intensity changes is the gradient of flattened patch. To obtain rotation invariance the idea of steerable filters [62] can be used. The Gaussian derivatives, on different orientations, are convolved with the extracted patch and all the outputs of the filters are represented as a vector. Shape context and spin images have also been used to represent partial object information using the boundary fragment model. The SIFT [10] is one of the most effective descriptors which

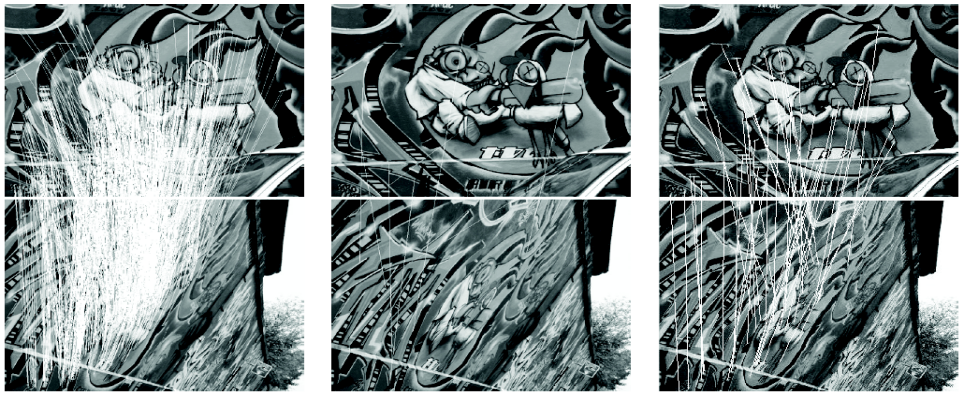


Figure 1.6: Performance comparison of ASIFT, SIFT and MSER for tilt value ≈ 3.2 [14].

divides a patch into 4×4 grid of cells followed by a histogram of oriented gradients for 8 orientation bins resulting into a vector with 128 dimensions. The gradient location and oriented histogram (GLOH) [113] and PCA-SIFT [118] are extensions of the SIFT. Recently, a affine version of SIFT has been proposed which permits reliable identification of features that have undergone very large affine distortions as measured by parameter called *transition tilt*. State-of-the art methods hardly

exceed transition tilts of 2 (SIFT) whereas ASIFT can handle transition tilts upto 36 and higher. Figure 1.6 shows the performance comparison of ASIFT against SIFT and MSER schemes (from left to right). The poor performance of Harris-affine and Hessian-affine have not been shown which found 3 and 1 correct matches for the similar image pair, respectively. Recently, Ferns a new patch corresponding scheme based on Naive Bayesian classifier has been proposed [85] and is capable of efficiently handling a large number of classes. The main idea is to recognize patches around point of interest by using multiple binary features to model posterior probabilities. The problem is computationally tractable by assuming independence amongst arbitrary set of features. The algorithmic implementation of the scheme does not require more than ten lines of codes and still its performance is remarkable on image sets containing very significant perspective changes. The performance comparison of matched rates

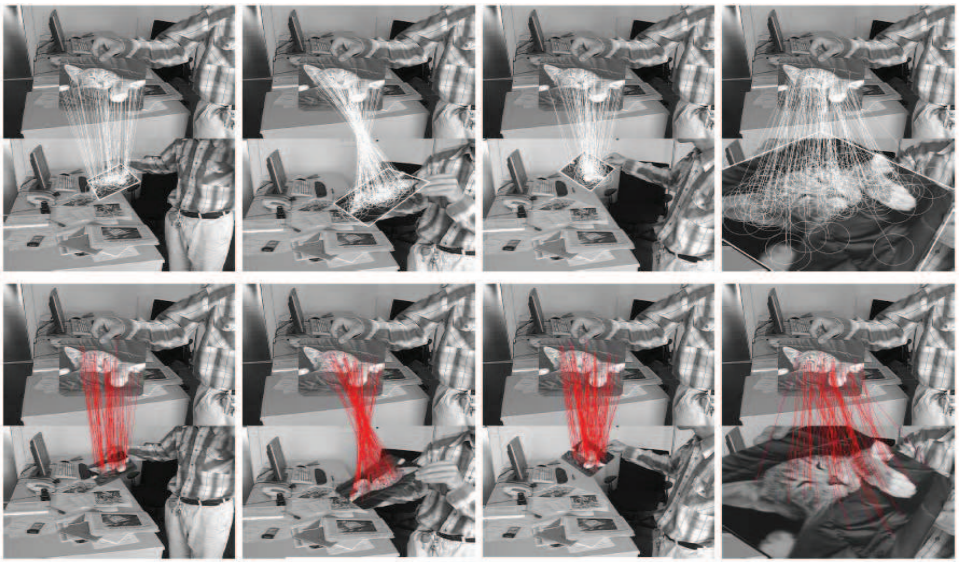


Figure 1.7: Detected matches for planar target using Ferns and SIFT (top to bottom) [85].

between SIFT and Ferns is presented in 1.7 where the top row shows matches obtained using Ferns and the bottom row is the representative performance for SIFT.

An image can also be represented as bag of features. In the training phase, a

database of patch features is created with the class of labels. For the testing phase, the detected patches are matched with the patches already stored in database. Since matching between patches can sometimes be accidently matched to the patches of objects from wrong categories, the content of an input image is labeled based on the majority voting scheme. For category level recognition, a histogram of the quantized patches, also called as *bag-of-visual-words*, is an effective representation. The quantization of local patches can be computed using clustering algorithms such as kNN. *Bag-of-visual-words* has profusely been used in recognition of objects and scenes for impressive results [47]. However, spatial relationship between features or their distribution is lost in the *bag-of-visual-words* representation of an image. The missing spatial layout is important to distinguish certain images which may have different objects but similar color histogram. To avoid such situation, color correlogram and probabilistic constellation of patches have been proposed [61],[74]. It should be noted that color correlogram computes the pair-wise relationship between two colors at a certain distance whereas the constellation model attempts to represent the probabilistic distribution of patch appearance and pair-wise patch correlation. In short, the constellation model describes the relative positions between pair-wise distinctive patches. In multiple scale matching instead of considering highest resolution of image partition, spatial pyramid match kernel [65] conducts image matching from the coarse to the finest level through histogram intersection. The matching is weighted by pyramid level since matching of different levels in a pyramid have varying importance. In general, the matches are performed at finer resolutions and are assigned higher weights whereas histogram intersection is used to measure the similarity between two models. Oliva and Torralba [66] presented a holistic approach by treating a scene as an object and attempted to describe the shape of a scene by a set of perceptual dimensions.

1.4.2 Background Survey of Action Recognition

From psychophysics points of view, it has been proven that humans are capable of recognizing actions solely from global information of the motion. Inspired by this research several techniques have been proposed to model human motion by capturing the global information using an attached marker. The use of an attached marker in applications like surveillance systems is infeasible because of practicality reasons as human body parts should be automatically detected for such problems. The vision based action recognition scheme can be mainly divided into two categories based on extracted features from a video sequence. The recognition belongs to the *holistic approach* if extracted features represent global shapes, contours and/or dimensionality reduced coefficients from feature space, otherwise, if it uses partial information of an incoming video it is termed *part based* approach.

Holistic Approach for Action Recognition

The correlation of a template action with an incoming video is considered to be the simplest holistic approach for action recognition. Features like intensity, gradient, and optical flow can be used in correlation [41]. The model based on correlation between optical flow field has been used for high action recognition rate [5]. The global information of a sequence of video frames can be described using silhouettes and contours. A 3D model of moving articulated subjects for multiple camera views is constructed using visual hull. The shape obtained from the reconstructed 3D model can be used to recognize human motion [42]. Instead of reconstructing a 3D model utilizing multiple cameras, the silhouettes can be directly embedded into a 3D space-time shape volume which supports a better representation of shape volume changes due to pose variations. Such integration of silhouettes results in transformation of an action recognition problem into 3D object recognition. The features extracted using differential geometry on the surface of the action volumes (refer to Figure 1.8) is proposed by Yilmaz and Shah with a reasonably good performance for action recognition [25]. To recognize an action from an arbitrary view, matching between 2D

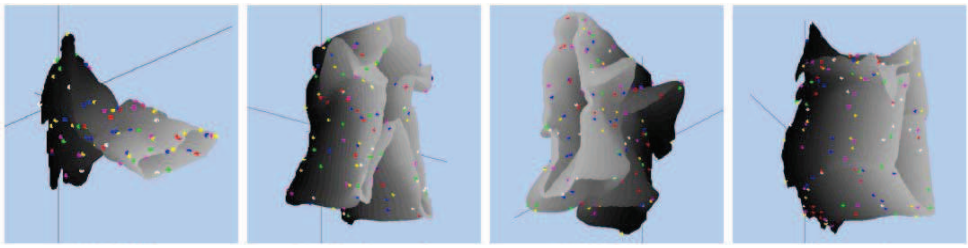


Figure 1.8: Sketches shown for actions of *falling, tennis stroke, walking and dancing* (left to right). Color codes are: red (peak), yellow (ridge), white (saddle ridge), blue (pit), pink (valley), green (saddle valley) [25].

silhouettes of the test sequence and key poses of each type of actions has been proposed [67] where such poses are also shared amongst various categories. The scheme to capture view independent motion changes by using silhouettes to generate motion history volume was proposed by Weinland [44] which is an extension of the idea to record the moving path of a motion into a single image [43]. The performance of silhouettes and shape matching algorithms solely depends on the results of background subtraction which is adversely affected by camera movements, lighting variations and occlusions. Therefore, matching solely a shape in global approaches is not reliable for highly accurate recognition frameworks. The use of multiple types of features for action recognition is getting increased attention in the research community due to the availability of complementary information which helps minimize the effects posed by a single feature type. Schneider proposed the combined use of motion and shape information detected by using optical flow and linear Gabor filters for action recognition [17]. Schneider's research shows impressive accuracy while using a lower number of past video frames, *snippets*, without the use of impractical assumptions such as the availability of look-ahead and foreground masks. The features from snippets are extracted in two parallel processing streams using different scales, orientations and speed while the filter responses are MAX-pooled and concatenated to be classified using a bank of linear classifiers. Although, this method was not designed for action recognition at the sequence level it achieves top performance on bench mark datasets

using bag-of-snippets of length one.

Part-based Approach for Action Recognition

Bag of features approaches have been proposed in the past to overcome the limitations of recognition schemes using global information such as background subtraction, tracking and variations in texture and color. Such features may include static and local features or the combination of both to represent an action. The human visual

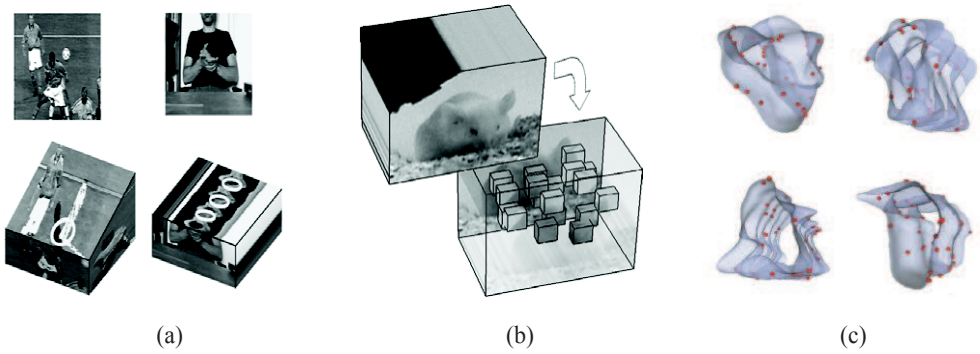


Figure 1.9: (a) Results of detecting the strongest STIP features (bottom row) for *football* and *clapping* sequence (top row) (modified version of [11]), (b) visualization of cuboid based behavior recognition [30], (c) sampled spin images for actions *bend, jack, walk* and *wave1* (red points are the oriented points) [12].

recognition is capable of recognizing an action from a single frame with cluttered background without necessarily requiring motion or temporal information. Such instantaneous postures, if selected precisely to provide important information, can also be used in the recognition process through extraction of features like HOG, appearance and position contexts [18],[68],[120]. Multiple frames can also be used instead of a single instantaneous pose in order to make up for the lacking information of motion, especially, in situations where motions features are not precisely extracted due to undesired variations in the scene or camera movements [47].

Since the introduction of the 3D Harris corner detector and the 1D Gabor filter in temporal direction to extract space-time interest points (STIP) and spatio-temporal (ST) features; these features have effectively been used in recognizing various motion

related tasks [30],[11],[35],[12],[13]. From Figure 1.9(a) it is clear that the identified points correspond to neighborhood with high spatio-temporal variations in sequence data. Recently, Ning *et al.* proposed a scheme based on the 3D Gabor filter to detect interest points to solve the problem of pose estimation. In [12] Liu *et al.* proposed the use of spin images (Figure 1.9(c)) which can provide a richer representation of changes of the local shape of an actor with respect to different reference points. These reference points may correspond to different limbs of the human body. Once, the spatio-temporal features have been detected then schemes like majority voting, statistical distribution or learning model can be used for classification [18],[45].

Combining multiple types of features is a relatively new field of research in action recognition. Fanti *et al.* [121] proposed the combination of velocity and local appearance descriptors. A generative scheme to learn the hierarchical model using both static and dynamic features has been proposed in [18] whereas Liu's work [12],[13],[47] and the results of Schindler's strategy [17] also verified the usefulness of integrating more than one types of features.

The rest of this dissertation is divided into four chapters. The next chapter presents a supervised recognition framework based on training performed analytically to avoid the traditional bottleneck of learning schemes. Chapters 3 and 4 are dedicated to fresh action recognition frameworks based on finder activity representation extracted using 3D dual-tree complex wavelet transform. The incremental learning, preserving neighborhood correlation information of video frames and real-time tracking based action recognition are noteworthy contributions in these chapters. At the end of this dissertation conclusion and future research plans are presented.

Chapter 2

Analytically Trained Recognition Framework

This chapter presents a new supervised learning scheme[1], which uses hybrid information, i.e., global and local features, for accurate identification and classification at a considerably high speed both in training and testing phases. The first contribution of this chapter is the unique image representation using bidirectional two-dimensional PCA and Ferns based approach to represent global and local information of an object, respectively. Secondly, the application of extreme learning machine supports reliable recognition with minimum error and learning speed significantly faster than traditional neural networks. The proposed method is capable of classifying various images in a fraction of second compared to other modern algorithms that require at least 2-3 seconds per image [87].

2.1 Object Recognition: Literature Survey

Object recognition or categorization is a task of classifying an individual object to belong to a certain category. Automated vision systems, in general, do not perform

[1]This chapter incorporates the outcome of a joint research undertaken in collaboration with A.A. Mohammed under the supervision of Dr. Q.M. Jonathan Wu [52].

better categorization than humans due to lack of intelligence, and knowledge. Despite some early success in automatic recognition; the problem is far from being solved due to preserved non-planar geometry and significant 3D depth variations in images of natural scenes. The image databases are an essential part of recognition research. For comparison of emerging algorithms; a number of publicly available databases have been established such as UIUC, Caltech, MIT, GRAZ and PASCAL. These databases provid a common ground for evaluation and assessment of fresh algorithms. Detecting objects in measurements is a complicated task owing to their enormously large number of possible poses, appearances in varying image acquisition conditions, and occlusions (see Figure 2.1 for sample images).

Computer vision community has been following a line of investigation to develop algorithms which can efficiently detect features, global or local, and regions for robust recognition of objects of interest. Each recognition method proposed in the past has its own merits and limitations; in general common approaches use image databases which contain object of interest at perceptible scale with minor deformations/pose variations. Feature extraction and representation of significant objects in an incoming image using the generated features is an initial step towards visual recognition. Next, a classifier is trained using the representation established at an earlier stage. The popular classifiers include support vector machines (SVM), Bayes classifier, Fisher linear discriminant and traditional neural networks, and hidden markov models (HMM) to name a few. For above classifiers; a degraded classification is observed due to non-convex feature space generated by the images captured under different geometric and lighting environment. Unfortunately, categorization turns out to be a complicated chore due to noticeable changes in appearance and other deformations caused by variations in the scene depth. Classification schemes use a wide variety of features like color, texture, orientation, blob, centre of gravity and mutual geometric relationship amongst feature points to learn a classifier. Visual recognition frameworks range from constellation of local features [74],[75], and complex geometric models [76] to the

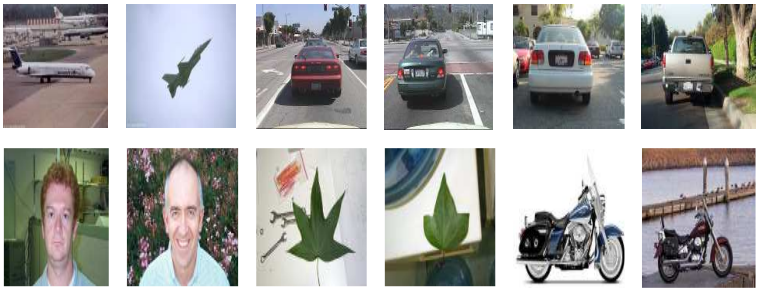


Figure 2.1: Sample images from Caltech database.

use of motion cues [77],[80]. Object categorization schemes with smaller variations in pose [74],[78],[79] and manual pre-segmentation of objects to minimize the computational cost have also been proposed [81]. Part based schemes [74],[78],[82],[86],[95] represent object structure using patches covering distinctive parts of an object. Such patches are extracted from neighborhood of interest points detected using localized operators like Harris corner detector. In [87], Ali and Shah proposed a promising approach to use the global structure of an object by modeling non-linear subspace of categories using Kernel PCA (KPCA) and selecting a discriminative feature set employing AdaBoost algorithm. Opelt et. al [86] used multiple kinds of features to encode the extracted patches and later used AdaBoost framework to select the best features for categorization.

The performance deterioration is observed in approaches which use only global information, especially in images with considerably large background clutter, geometric deformations, and occlusions. The correlation amongst neighborhood pixels is also ignored due to vectorization of an image during dimensionality reduction operations such as PCA and kernel PCA. On other hand, part based recognition schemes are computationally expensive requiring significantly large amount of training samples (extracted/synthesized for different view points) that eventually leads to momentous increase in computational cost. Above precincts lead to two important but overlooked problems in the past recognition studies.

- **How to minimize the training time** which poses a major bottleneck in classification.

- **How to efficiently retain the correlation information amongst neighboring pixels** which is lost during customary dimensionality reduction operations such as PCA and KPCA.

We propose a hybrid approach for recognition that combines global and local object information for robust and reliable recognition. The use of two-dimensional PCA (2D-PCA) [92] along mutually orthogonal directions is proposed to encode global information of an image which can better preserve association amongst neighboring pixels. Technique based on multidimensional PCA [93] is recently proposed for face recognition. The use of general tensor discriminant analysis (GTDA) for gait recognition, proposed by Tao et. al [99], is proven to generate improved accuracy with minimal undersample problem during classification. In [100],[101] incremental and supervised approaches for tensor analysis have been proposed which generate better-quality recognition with structure preserving processing in higher dimension data similar to bidirectional 2D-PCA. However, our recognition method differs from [99] in two ways 1) using hybrid object information for classification 2) synthesizing images for various affine deformations to train our classifiers for potential objects' views. For local object information, feature vectors are generated from patches around stable feature points detected using Harris corner detector. Multiple views of such patches are generated through affine deformations which result into considerably increased number of training samples. Later, extreme learning machine (ELM) [88] is applied for recognition using both kinds of feature vectors i.e. global and local information.

The use of any other supervised learning framework such as neural network or AdaBoost may require longer training intervals due to their specific learning strategy while ELM can finish the similar training task at speed approximately thousands times faster than traditional neural networks and minimum training error. ELM has

been successfully applied for multicategory classification such as microarray gene expression for cancer diagnosis [89] and classification of music genres [90]. Our proposed method allows to combine the strengths of both types of features and exploits highly discriminative feature sets for classification using ELM. A wide variety of experiments using standard datasets are presented to ascertain the superior performance of our proposed scheme over other state-of-the-art methods.

2.2 Extreme Learning Machine

Feedforward neural networks (FNN) have been widely used in different areas due to their approximation capabilities for non-linear mappings using input samples. It is a well known fact that the 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. [88] 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 Figure 2.2). 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 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 strategy of traditional feedforward neural networks like back propa-

gation (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

$$\sum_{i=1}^{L} \beta_i g(x_l) = \sum_{i=1}^{L} \beta_i g(w_i.x_l + b_i) = o_l, \ l \in \{1, 2, 3, \ldots, N\} \tag{2.1}$$

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 the inner product of w_i and x_l. Above modeled ELM can reliably approximate N samples with zero error as

$$\sum_{l=1}^{N} \|o_l - \gamma_l\| \ = \ 0 \tag{2.2}$$

$$\sum_{i=1}^{L} \beta_i g(w_i.x_l + b_i) \ = \ \gamma_l, \ l \in \{1, 2, \ldots, N\}. \tag{2.3}$$

where 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 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,

$$\Upsilon = (w_1, \ldots, w_L, b_1, \ldots, b_L, x_1, \ldots, x_N). \tag{2.4}$$

$$\Upsilon = \begin{bmatrix} g(w_1.x_1 + b_1) & \cdots & g(w_L.x_1 + b_L) \\ \vdots & \cdots & \vdots \\ g(w_1.x_N + b_1) & \cdots & g(w_L.x_N + b_L) \end{bmatrix}_{N\times L} \tag{2.5}$$

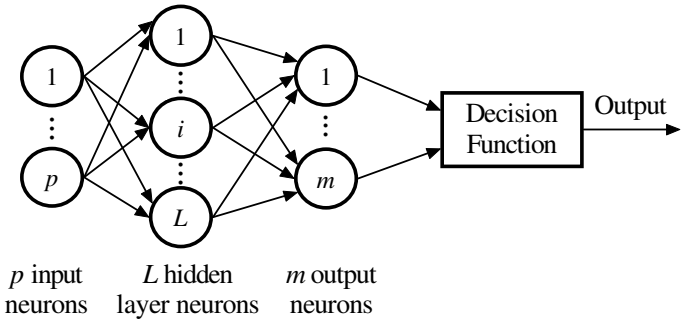


Figure 2.2: Simplified structure of ELM.

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

$$\varepsilon = \sum_{l=1}^{N}(\sum_{i=1}^{L} \beta_i g(w_i x_l + b_i) - \gamma_l)^2 \tag{2.6}$$

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 is tuned iteratively by

$$w_t = w_{t-1} - \rho\frac{\partial\varepsilon(W)}{\partial W}. \tag{2.7}$$

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. [88] 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 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

$$\|\Upsilon(\hat{w_1}, \ldots, \hat{w_L}, \hat{b_1}, \ldots, \hat{b_L})\hat{\beta} - \Gamma\| \tag{2.8}$$

$$= \min_{w_i, b_i, \beta} \|\Upsilon(w_1, \ldots, w_L, b_1, \ldots, b_L)\beta - \Gamma\|. \tag{2.9}$$

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

$$\|\Upsilon\hat{\beta} - \Gamma\| = \|\Upsilon\Upsilon^*\Gamma - \Gamma\| \tag{2.10}$$

$$= \min_{\beta} \|\Upsilon\beta - \Gamma\|. \tag{2.11}$$

2.3 Feature Extraction

A systematic framework for object recognition is presented based on hybrid object information. Successful extraction of good features from images is crucial to object recognition considering large variations in realistic images.

2.3.1 Global Feature Vector Computation

Karlhunen-Loeve expansion, also known as principal component analysis (PCA), is a data representation technique widely used in pattern recognition and compression schemes. In [92], Yang et al. proposed two-dimensional PCA for image representation. 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 denote an M-dimensional unitary column vector. To project a $Q \times M$ 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 data 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., $J(X) = \mathtt{tr}(S_x)$ where $\mathtt{tr}(.)$ 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

$$\begin{aligned} S_x &= E[(Y - E(Y))(Y - E(Y))'] && (2.12) \\ &= E[[(A - E(A))X][(A - E(A))X]'] && (2.13) \\ \mathtt{tr}(S_x) &= X'[E(A - E(A))'(A - E(A))]X. && (2.14) \end{aligned}$$

The image covariance matrix is defined as $G_t = [(A - E(A))'(A - E(A))]$. It is easy to verify that G_t is a $M \times M$ nonnegative definite matrix; suppose that there are P training image samples, the j^{th} sample of size $Q \times M$ is denoted by A_j where $1 \leq j \leq P$. G_t is computed by

$$\begin{aligned} G_t &= \frac{1}{P}\sum_{j=1}^{P}[(A_j - \overline{A})'(A_j - \overline{A})] && (2.15) \\ J(X) &= X'G_tX && (2.16) \end{aligned}$$

where $\overline{A}$ represents the average image of all training samples. Above criterion is

Table 2.1: Time spent for training and testing using ELM on global feature vectors.

Time	Planes		Background		Cars		Bikes		Faces		Leaves	
	TT	CT	TT	CT	TT	CT	TT	CT	TT	CT	TT	CT
Planes	N/A	N/A	.078	.062	.140	.062	.078	.078	.078	.062	.094	.047
Background	.094	.078	N/A	N/A	.047	.047	.094	.047	.109	.031	.156	.016
Cars	.156	.094	.031	.047	N/A	N/A	.125	.047	.140	.031	.125	.031
Bikes	.125	.078	.078	.047	.094	.047	N/A	N/A	.109	.047	.078	.047
Faces	.094	.047	.140	.031	.109	.031	.094	.047	N/A	N/A	.109	.016
Leaves	.094	.047	.109	.016	.125	.016	.140	.047	.094	.016	N/A	N/A
TT: Training Time (sec.), **CT**: Classification Time (sec.)												

called the *generalized total scatter criterion*. The unitary vector X that maximizes the criterion is called the optimal projection axis. We usually are required to select a set of projection axes, $X_1, X_2, \ldots, X_d$ (where subscript d is a scalar value representing the number of dimensions), subject to orthonormal constraint and to maximize the criterion $J(X)$. Yang et al. [92] showed that extraction of image features using 2D-PCA is computationally efficient and better recognition accuracy is achieved than 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. Zhang and Zhou [94] proposed $(2D)^2$ PCA based on assumption that training sample images are zero mean and image covariance matrix can be computed from the outer product of row/column vectors of images. We propose a modified bidirectional 2D-PCA to extract features 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 zero. The vectorization of mutual product of such covariance matrices results into a considerably smaller sized feature vectors which retain better structural and correlation information amongst neighboring pixels. Figure 2.3 shows better ability of bidirectional 2D-PCA to represent the global structure of various object categories. Figures 2.3(a-b) are plotted using Caltech (Airplanes and Leaves) and MIT (Cars and Pedestrians) datasets respectively, whereas Caltech (Airplanes and Motorbikes) datasets have been used for Figures 2.3(c-d). The first two components of feature vectors obtained using bidirectional 2D-PCA and Kernel PCA are plotted against each other. In Figures 2.3(a-b); we observe nearly separated classes for Caltech and MIT datasets; this validates our claim that 2D PCA achieves superior categorization (see Table 2.2) for these datasets. For kernel PCA (Figure 2.3(d)); it is quite clear that the first two components, representing the largest eigenvalues, are almost identical and analogous overlap of these feature vectors may lead to poor classification. For the same datasets; use of bidirectional 2D-PCA generates classes which are partly converged as shown in Figure 2.3(c). Table 2.1 demonstrates the

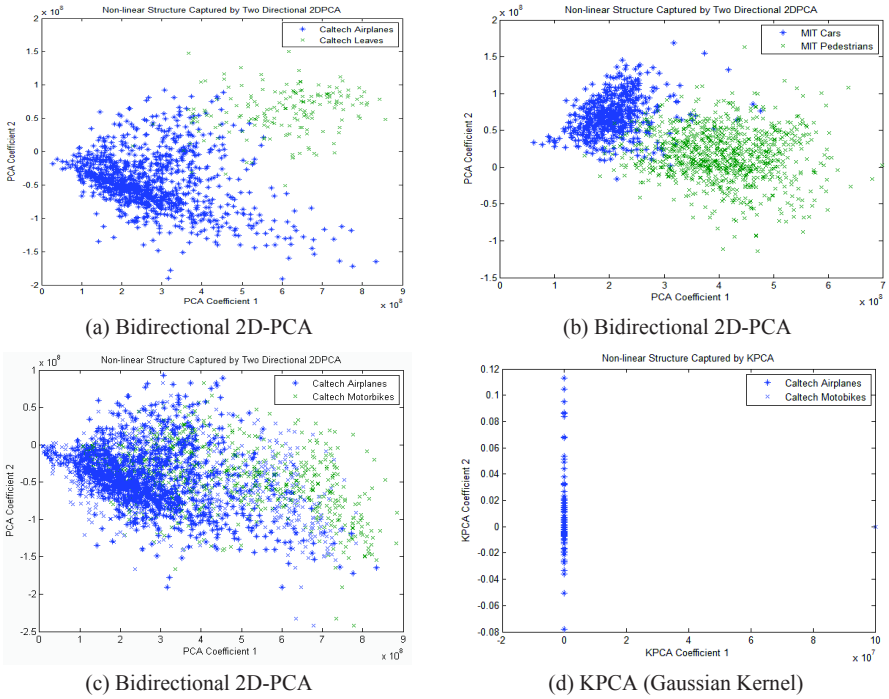
Table 2.2: Training and detection accuracy using ELM on global feature vectors.

Accuracy	Planes		Background		Cars		Bikes		Faces		Leaves	
	TA	CA	TA	CA	TA	CA	TA	CA	TA	CA	TA	CA
Planes	N/A	N/A	98	86	100	100	99.5	91.2	100	90.6	100	97.6
Background	96	79.8	N/A	N/A	100	100	97.5	73.7	99.5	82.6	99	92
Cars	100	100	100	100	N/A	N/A	100	100	100	100	100	100
Bikes	99.5	91.6	99	80.1	100	100	N/A	N/A	100	93.8	100	95
Faces	100	97.1	99.5	87	100	100	100	91.3	N/A	N/A	100	92.7
Leaves	100	97.2	97.5	85.8	100	100	100	96.8	100	95.4	N/A	N/A
TA: Training Accuracy (%age), **CA**: Classification Accuracy (%age)												

extremely fast classification capability of ELM using global features. However, the accuracy achieved using these global features is not stable and varies with the selection of datasets in different combinations to represent positive and negative classes during recognition (see Table 2.2 for mutual combinations of Airplanes, Background and Motorbikes from Caltech datasets). Therefore, we propose to combine complementary information, i.e. local feature vectors, along with the global contents of an image to attain reliable classification which is independent of view point changes and dataset combinations. It should be noted that all Matlab implementations of our experiments are executed on a desktop computer equipped with Intel Core 2 Duo processor of 2.6 GHz speed and 2GB RAM.

2.3.2 Local Feature Vector Computation

Identifying textured patches that are distinctive and detectable under varying pose and lighting conditions in neighborhood of stable feature points is a widely researched area with numerous applications. Different strategies have been proposed to use local patches or contour based information for object detection with features being shared



(a) Bidirectional 2D-PCA (b) Bidirectional 2D-PCA

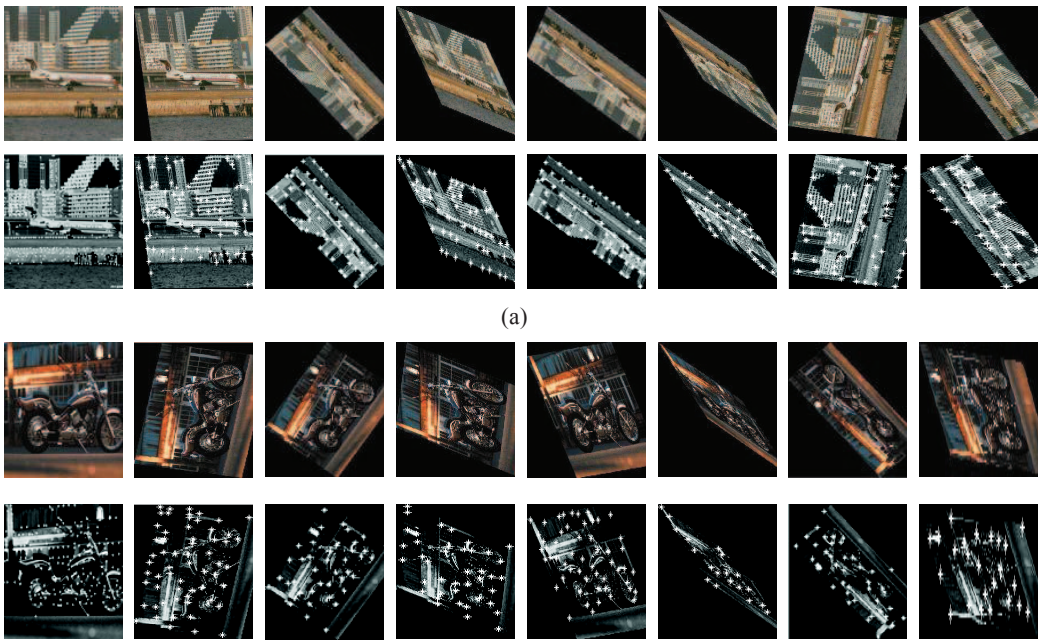
(c) Bidirectional 2D-PCA (d) KPCA (Gaussian Kernel)

Figure 2.3: Non-linearity capture among datasets.

among different classes [77],[95]-[98]. This section proposed extraction of partial object information using feature vectors computed from local patches surrounding a set of stable feature points. These feature vectors are used as complementary information to enhance classification accuracy.

A semi-naive based classifier, recently proposed by Mustafa et al. [85], is used to determine the class of local patches surrounding stable key points. The solution for patch correspondence problem provided in [85] shows promising results, comparable to state-of-the-art, yet simple by exploiting statistical information of pixel intensities. To detect preliminary stable key points, randomly selected images of a specific class are deformed and Harris corner detector is applied. We select Harris corner detector for its simplified and efficient implementation to detect key points with minimal computational burden compared to other schemes such as SIFT, complex filters, PCA-SIFT, and cross-correlation [83],[84],[102]. The parameters for affine

deformations are randomly picked from a uniform distribution therefore two images of a similar object may have differently been warped at two different time instances. The corners identified for deformed set of images change based on chosen parameters and background clutter. It is realized that the rising number of affine warped images can lead to higher computational load. However, comprehensive training sets to mimic possible appearances of local patches generate improved recognition. On other hand, such a pronounced computational complexity is defied only in training whereas the recognition of an incoming image during testing phase is undemanding and high-speed procedure. Figure 2.4 shows different feature points detected at vary-



(a)

(b)

Figure 2.4: Preliminary key points detected under varying poses.

ing deformations; two sample images from Caltech airplane and motorbikes datasets are used and it is obvious that number of detected feature points are changing in different transformations. A list is maintained to keep track of points, which have been repeated the most, to extract local patches. We declare the identified key points as *stable* if the their rate of recognition is more than 75% of the total number of

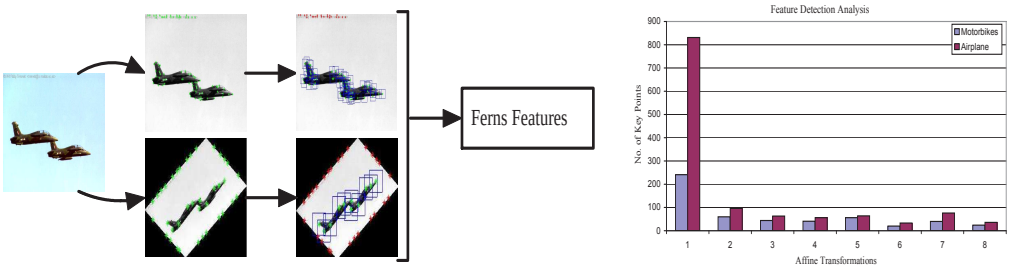


Figure 2.5: *Left*: Extracted local patches for Ferns computation, *right*: Histogram of identified key points for varying affine deformations.

deformations. A high variation in number of detected feature points is observed; please refer to histogram presented in Figure 2.5 where the number of identified key points is significantly changing for varying affine parameters. For practicality, white noise is also added so that the patch is processed in conditions akin to a real life situation. The patches surrounding stable key points of size 16×16 are extracted whereas the deformed versions of images help to achieve synthesis to symbolize the possible appearances under varying poses. Figure 2.5(left) shows extracted patches (blue squares), with a height and width of 16 pixels each, using an illustration image from Caltech (Airplanes) dataset; please note that the recognized stable key points are represented by the green color whereas outliers have been represented by red. After extraction of local patches; assigning each patch to a most probable object class is a subsequent task. Let $c_i, i = 1, 2, \ldots, H$ be a set of classes and $f_j, j = 1, 2, \ldots, Z$ be a set of binary features to be computed from extracted patches. We want to classify a patch based upon binary features as follows:

$$\hat{c}_i = \arg\max_{c_i} P(C = c_i | f_1, f_2, \ldots, f_z), \tag{2.17}$$

$$P(C = c_i | f_1, f_2, \ldots, f_z) = \frac{P(f_1, f_2, \ldots, f_z | C = c_i) P(C = c_i)}{P(f_1, f_2, \ldots, f_z)}. \tag{2.18}$$

Assuming uniform prior probability $P(C)$ and denominator $P(f_1, f_2, \ldots, f_z)$ as scaling factor; our problem is reduced to

$$\hat{c}_i = \arg\max_{c_i} P(f_1, f_2, \ldots, f_z | C = c_i). \tag{2.19}$$

The computation of each binary feature f_j depends upon mutual relationship of two pixel intensities located at $d_{j,1}$ and $d_{j,2}$ in the patch.

$$f_j = \begin{cases} 1, \text{ if } I(d_{j,1}) < I(d_{j,2}) \\ 0, \ otherwise \end{cases} \tag{2.20}$$

where $I(.)$ represents an image patch. Assuming a complete independence between features leads us to

$$P(f_1, f_2, \ldots, f_z | C = c_i) = \prod_{j=1}^{z} P(f_j | C = c_i). \tag{2.21}$$

However, the correlation amongst neighboring pixels of a patch is ignored hence an acceptable compromise can be modeled as

$$P(f_1, f_2, \ldots, f_z | C = c_i) = \prod_{k=1}^{M} P(F_k | C = c_i) \tag{2.22}$$

where M represents the number of feature clusters of size $S = Z/M$ each, a Fern F_k is represented by

$$F_k = f_{\sigma(k,1)}, f_{\sigma(k,2)}, \ldots, f_{\sigma(k,S)} \tag{2.23}$$

where $\sigma(k, S)$ shows a random permutation function with range $1, \ldots, Z$. A reliable and fast patch correspondence using above relationship is reported in [85]. A performance and computational load trade-off is observed for varying values of M and S. In training phase, class condition probabilities for individual ferns are estimated which are combined to label corresponding extracted patches. We generate $M + 1$ dimensional local feature vectors for individual patches which comprise of conditional probabilities of mutually independent ferns and their combined information to compute the conditional probability of a local patch. Finally, such feature vectors are used for training and testing purposes using ELM as classifier working on local type of features.

2.4 Learning Classifiers with Parallel ELMs

The classification algorithm is provided with a set of training images where positive label indicates that an object of interest is present in an image while negative label represents its absence. All images are converted to gray level and resized to square dimension matrices. There is no further pre-processing applied to datasets and we assume no prior information about location, view point and/or image acquisition constraints. To avoid the curse of dimensionality, bidirectional 2D-PCA is employed which requires multiplication between two covariance matrices (one for each of row and column directions). The output of this dimensionality reduction step is a square matrix which is vectorized and termed as global feature vector.

The proposed recognition scheme declares an incoming image as positive class if the relevant object is present. For fast pre-processing direct intensity values are used to extract both kinds of features, i.e., global and local. Generating a variety of features representing various contents of an image allows to knob varying geometric attributes of an object and achieve better categorization. Figure 2.6 represents a generalized framework that supports integration of a wide variety of learnable local descriptors for enhanced classification. We used only single type of local feature vector generated using Ferns [85] style patches surrounding stable feature points identified using Harris corner detector. Due to significantly shorter training time, and mini-

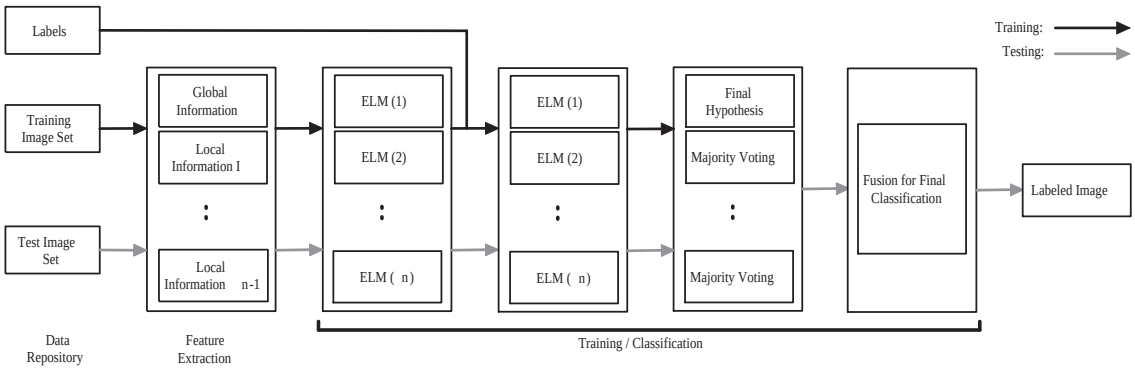


Figure 2.6: Different steps involved in our algorithm.

mized computational burden; we use more than one ELMs in a parallel fashion to process all categories of image feature simultaneously for real-time classification. The training process for an ELM operating on global feature vectors is not the same as one using set of local features. Computed global training feature vectors are directly input to an ELM, whereas training for a ELM that deals with local features, starts with application of corner detector by deforming the training images and keeping a track of the number of times same feature point is identified. Such image deformations are suggested to train our classifier for possible pose variations, and is proved to be feasible due to fundamentally soaring training speed (see Tables 2.1, 2.3). The number of feature vectors representing the local patches of an image may vary depending upon stable key points detected from a synthesized set of images for different affine deformations. A majority voting scheme is adopted for reliable estimation of an image class since we observe false alarms for individual local patches due to low information content and an accidental matching among different regions of two different objects. The ELM operating on global feature vectors does not require voting scheme since *one-to-one* correspondence holds between feature vectors and individual images. Finally, a fusion process is initiated to combine n estimates originating from all ELMs based on normalized weighted sum strategy that allows us to assign importance to each approximation based on confidence (as follows):

$$Label \;=\; \begin{cases} +1, \text{ for } \sum_{i=1}^{n} w_i.e_i \geq Th; \sum_{i=1}^{n} w_i = 1 \\ -1, \; otherwise \end{cases} \tag{2.24}$$

where w_i,e_i and Th represent weight, estimate for individual ELM and threshold respectively. In our proposed framework, user has better control over preference to be given to an individual feature type. Since the penchant strategy for different feature types is solely dependent upon application, photometric and geometric elements of individual objects. The value for threshold, i.e., Th may vary between zero and one depending upon required confidence. It is obvious that a high value of Th may

result into increased reliability of classification with lower false positives and increased chance of false negative alarms. During experiments, we assigned a 0.5 weight values to each of the estimates originating from two ELMs operating on both kinds of feature vectors whereas a threshold of 0.75 is setup for final classification. Different steps involved in our proposed algorithm are presented in Figure 2.6.

Table 2.3: computational time (sec.) for GRAZ dataset.

		Bikes	Cars	Persons
Training Time	Bikes	N/A	4.29	4.31
	Cars	4.30	N/A	4.34
	Persons	4.30	4.37	N/A
Classification Time	Bikes	N/A	3.21	2.93
	Cars	3.25	N/A	3.01
	Persons	2.91	2.96	N/A

2.5 Results and Discussion

We used standard datasets to test the viability of our proposed method. The datasets from Caltech include Airplanes, Cars Brad, Faces, Leaves, Background, and Motorbikes whereas GRAZ and MIT image sets comprised of Bikes, Cars, Persons, and Pedestrians respectively. Table 2.3 presents the CPU time allocated for classification of GRAZ datasets using 53712, 53455, and 46495 features for Bikes, Cars and Persons datasets respectively. It is clear from the allocated CPU time that our proposed algorithm performs categorization at tremendously swift speed. In addition to speed, accuracy is another critical issue to judge the competence of a classifier. Numbers of experiments using challenging categorization image sets have been performed to analyze the performance of our ELM based classifier; the classification results of MIT and Caltech datasets are presented here for comparison of classification accuracy. Figure

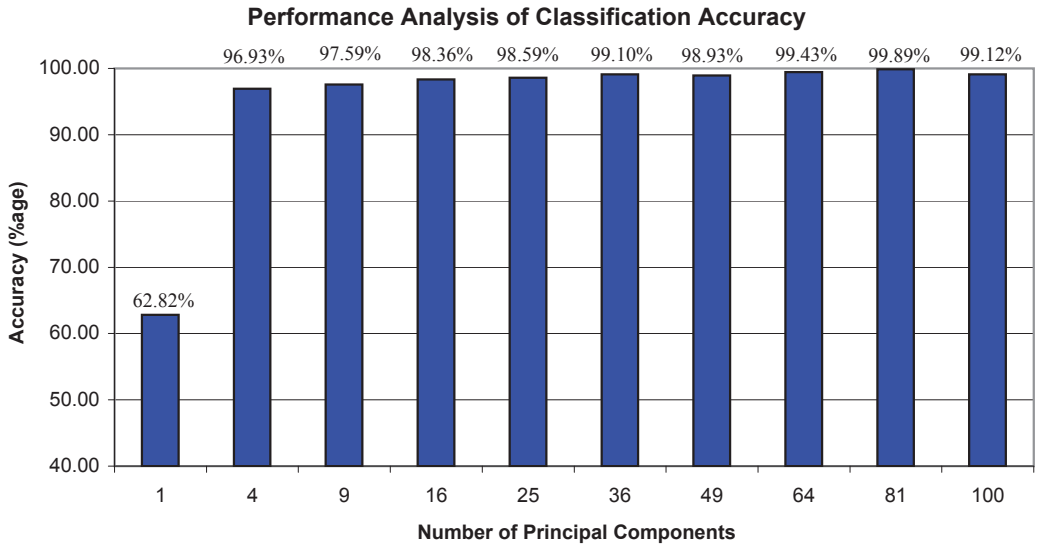


Figure 2.7: Classification accuracy using MIT dataset for varying number of principal components.

2.7 shows accuracy achieved for classification of MIT datasets (Cars and Pedestrians used as positive and negative classes respectively). Above 95% categorization accuracy is achieved for MIT database using multiple principal components, it is also realized that increasing the number of principal components is not a solution to improve detection accuracy. Our experiments are based on binary classification problem however they can be extended to multiclass scenario with minor modifications. It is an interesting aspect to investigate the impact of changing threshold Th on accurate classification. The optimal value for Th can help to minimize false alarms and precisely identify the object class present in an input image. Since we do not get classification with perfect confidence because of noisy measurements and various distorting parameters during image acquisition therefore we try to estimate an acceptable compromise between accuracy and confidence. The adjustment of Th to a value of 1 may lead to rejection of correct classification with lower confidence and increased false alarm ringing is witnessed for lower values of threshold. Various experiments are conducted with changing values of Th to obtain an optimal solution, the number

Table 2.4: Details of datasets used in analysis of changing threshold against accuracy.

	Positive Class	Negative Class	Training Images	Testing Images
Caltech	Leaves	Faces	200	436
MIT	Pedestrians	Cars	300	1140
GRAZ	Bikes	Persons	200	476

of principal components to represent global feature vectors are fixed to 36 whereas number of training and test images are different for specific datasets based on their respective sizes. For these trials, we randomly picked Leaves and Faces, Pedestrians and Cars, Bikes and Persons to represent positive and negative classes from Caltech, MIT and GRAZ datasets respectively. The number of images for both classes of objects also varies with datasets and readers may refer to Table 2.4 for further details. The achieved accuracy for various image sets is represented in Figure 2.8 for varying

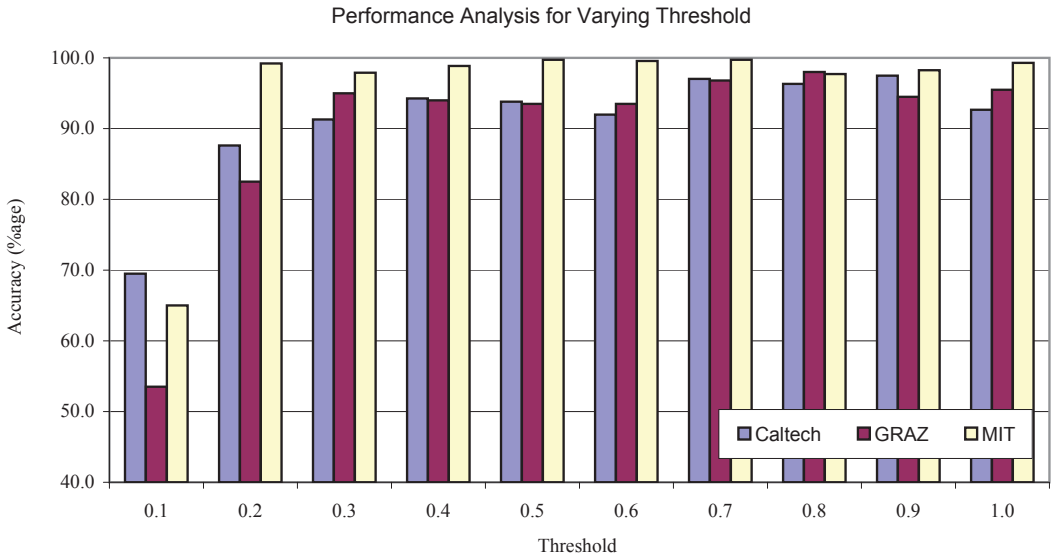


Figure 2.8: Performance analysis for changing threshold.

values of threshold Th. The proposed method performs considerably well for MIT dataset on changing threshold however an inconsistent classification is observed for rest of the datasets included in our experiments. One may notice that the devia-

tion in achieved correctness is higher for the combined results of all three input sets. However, a decrease in deviation amongst generated classifications is observed for $0.7 \leq Th \leq 0.8$. Such smaller deviation value amongst classifications for all datasets represents the steady performance of our scheme for images with differing geometric and radiometric variations. Therefore, one may conclude, based on experimental evidence, that the value of Th set to a lower value or very close to 1 may lead to rising false alarms and rejection of correctly classified object with lower confidence due to uncertain conditions. The Caltech dataset is used to test the performance of our proposed framework against state-of-the-art. The images are randomly selected from various Caltech datasets to represent negative class. For Caltech datasets; Table 2.5 presents an accuracy comparison of categorization for different modern algorithms; our proposed method achieves an average accuracy above 97% and outperformed other well-established schemes.

Table 2.5: Accuracy comparison for different approaches (%).

Dataset(s)	Our Method	[87]	[97]	[74]
Bikes	94.6	93.4	92.5	73.9
Planes	95.3	90.0	90.2	92.7
Cars	99.0	96.0	90.3	97.0
Leaves	98.3	94.2	-	97.8
Faces	97.9	98.0	96.4	-

2.6 Summary

We present a novel supervised learning algorithm for object detection and categorization that combines the strengths of both global and local features and demonstrate its considerably high speed in both training and testing phases. The proposed framework is capable of handling changes in pose, illumination, inter-class and intra-class

attributes. The proposed parallel architecture where each ELM module is simultaneously working, on distinct feature types, formulates a potential classifier for problems requiring significantly faster and reliable categorization. Features obtained through synthesized views of extracted local patches add further information to classification which is partially invariant to pose and lighting conditions.

Chapter 3

Visual Vocabularies for Human Action Recognition

This chapter introduces a novel recognition framework for human actions using hybrid features[1]. 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 '3D box of data' instead of frame by frame analysis, 2) rich representation of human actions with reduced artifacts using completely symmetrical complex filter banks. No assumptions of scene background, location, objects of interest, or point of view information is made. Bidirectional two-dimensional PCA (2D-PCA) is employed for dimensionality reduction as it preserves structure and correlation amongst neighborhood pixels of a video frame.

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] (large number of features may result in higher computational load). Generally, the precision of optical flow estimation is reliant upon tribu-

[1]This chapter incorporates the outcome of a joint research undertaken in collaboration with A. Baradarani, S. Seifzadeh under the supervision of Dr. Q.M. Jonathan Wu [54, 55, 105]

lations in aperture and properties of the surface being captured. Geometrical model [1, 8, 15] of local human parts is used to recognize the action using static stances in a video sequence that 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 action classification. In [6] promising results are achieved assuming that background is known for preliminary segmentation. Space-time interest points for action recognition have been proved to be a thriving technique [11, 27, 29, 30, 32, 35] independent of pre-segmentation or tracking of individual dynamic objects in a video. To improve classification performance, both shape and spatio-temporal features have also been 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 a 3D filter that consumes higher computational time. The ST volumes around interest points are extracted for further processing. To detect events in a video sequence, the extraction of STIP features is based on the idea of Harris and Förstner interest point operators. It is 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 chapter, 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 is constructed by designing an appropriate pair of orthogonal or biorthogonal filter banks that work in parallel. 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 [4] and further developed as 3D wavelet transform in video denoising by Selesnick et al. [19, 20]. This 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 as input to a classifier in order to recognize an action present in an arriving video. Extreme

learning machine (ELM) is a supervised learning framework [88], single hidden layer feedforward neural network, that is trained thousands times faster speed than traditional learning schemes such as gradient descent. ELM is applied to classify the actions represented by visual vocabularies.

3.1 Preliminaries

3.1.1 Dual-tree Complex Filter Banks

Consider the two-channel dual-tree filter bank implementation of the complex wavelet transform as shown in Figure 3.1. The primal filter bank $\mathbf{B}$ in each level defines the real part of the wavelet transform and the dual filter bank $\widetilde{\mathbf{B}}$ represents 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 $\mathbf{B}$ 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 ($\widetilde{\phi}_h$ 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

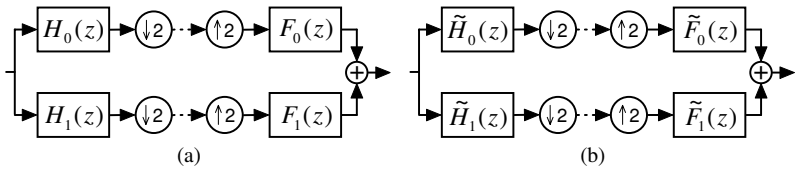


Figure 3.1: (a) The primal filter bank $\mathbf{B}$; (b) The dual filter bank $\widetilde{\mathbf{B}}$.

$\widetilde{\psi}_f + j\psi_f$, if the wavelet functions of the two filter banks form an Hilbert transform pair. Specifically, the analysis wavelet $\widetilde{\psi}_h(t)$ of $\widetilde{\mathbf{B}}$ 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 $\widetilde{\psi}_f(t)$. 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 [37]. This introduces limited redundancy and allows the transform to provide approximate shift-invariance and more directionality selection of filters [9, 21]. It preserves the property of perfect reconstruction and achieves 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

$$H_0(\omega)F_0(\omega) + H_1(\omega)F_1(\omega) = 1 \tag{3.1}$$

and the no-aliasing condition

$$H_0(\omega + \pi)F_0(\omega) + H_1(\omega + \pi)F_1(\omega) = 0. \tag{3.2}$$

The above no-aliasing condition is automatically satisfied if

$$H_1(z) = F_0(-z) \text{ and } F_1(z) = -H_0(-z). \tag{3.3}$$

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

$$\widetilde{H}_1(z) = \widetilde{F}_0(-z) \text{ and } \widetilde{F}_1(z) = -\widetilde{H}_0(-z). \tag{3.4}$$

where z refers to the z-transform.

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

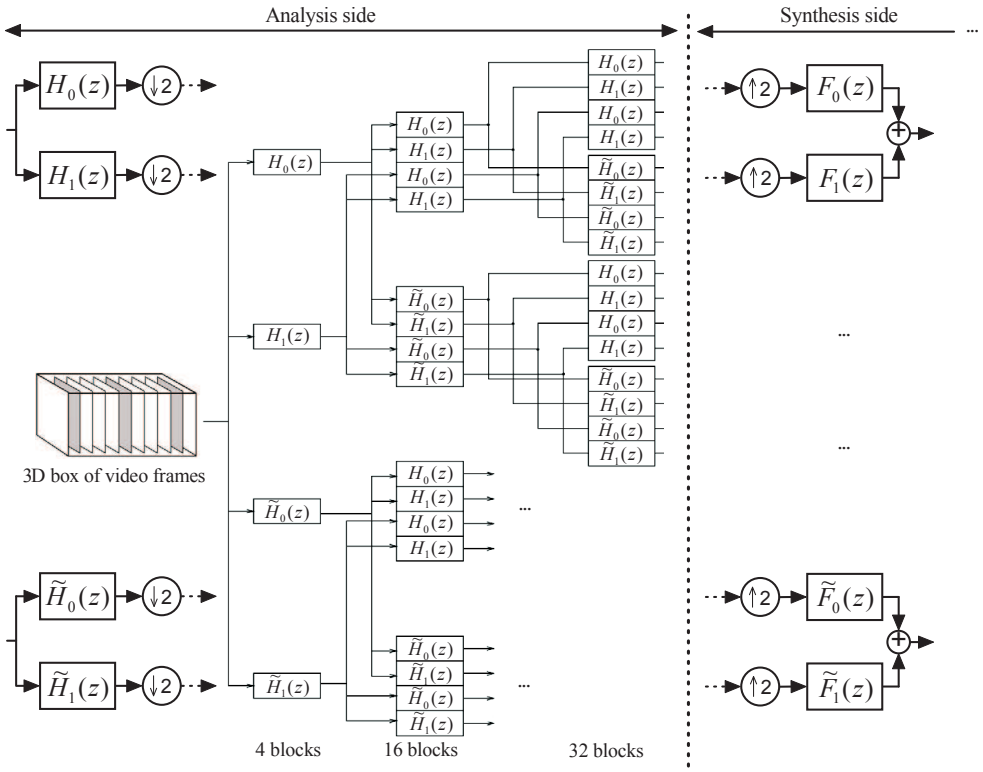


Figure 3.2: Typical schematic of filters in a 3D DT-CWT structure with the real and imaginary parts of a complex wavelet transform. 28 of the 32 subbands are wavelets excluding the scaling terms. Only the analysis side is shown in this figure.

wavelet transform. This is related to the real and imaginary parts of a 3D complex wavelet with two wavelets in each direction. Figure 3.2 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 better localize motion in its several checkerboard-free directional subbands compared with 2D-DWT and separable 3D-DWT. 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.

3.2 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 is 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. In this work we do not assume any *a priori* information about background, view point, activity and data acquisition constraints.

3.2.1 Synopsis of Proposed Framework

The implementation of proposed algorithm starts with the computation of hybrid feature vectors. As preprocessing operation, incoming video frames are converted to gray space and resized to square dimension. 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 multires-

olution 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 as input to ELM. 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 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 Figure 3.3.

3.2.2 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 without any segmentation and stabilization operation. This is an important contribution, the techniques proposed in the past assumed knowledge of background or foreground masks or required manual stabilization operation of 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 similar or dissimilar actors. Applying 3D DT-CWT on an input video sequence of

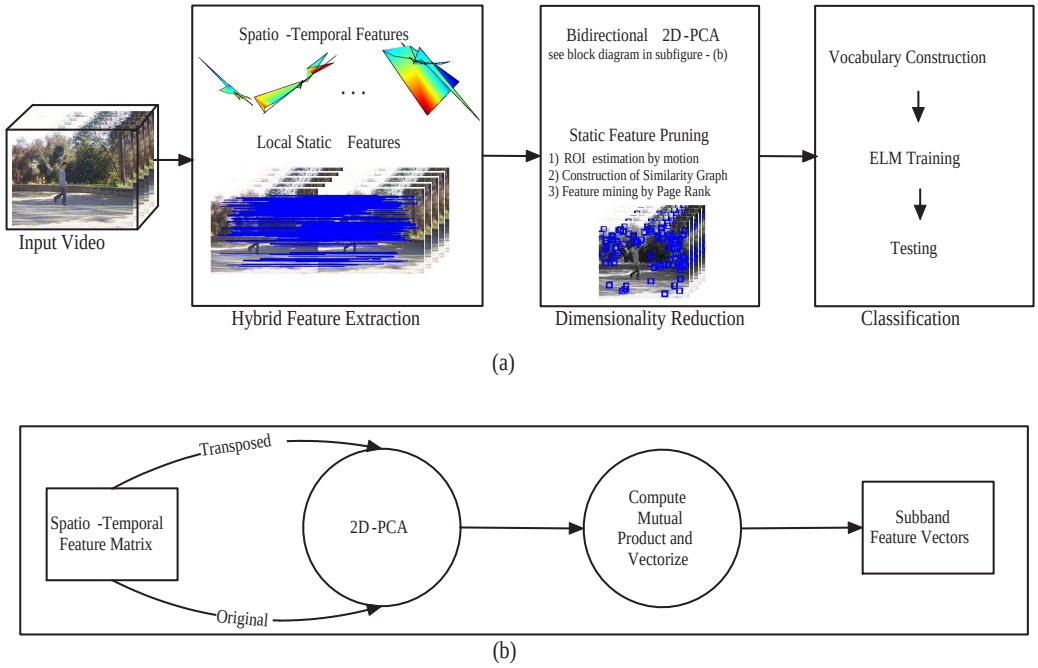


Figure 3.3: The block diagram of our proposed algorithm (a) main steps of the proposed scheme (b) steps involved in computation of bidirectional 2D-PCA.

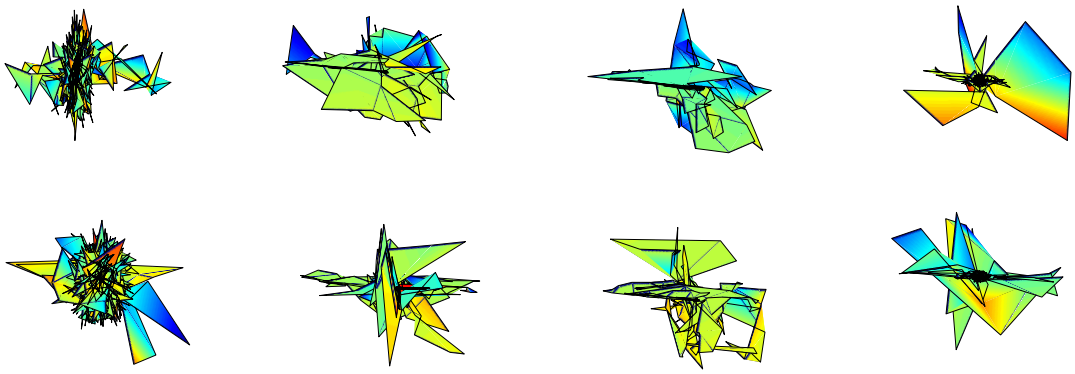
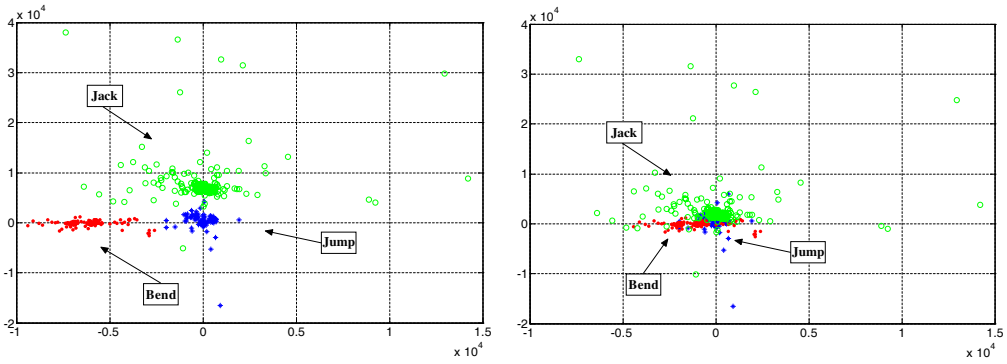


Figure 3.4: Some sample spatio-temporal features computed using motion selectivity attribute of 3D DT-CWT. From left to right columns, top view of first directional subband for four actions, namely, *bend*, *run*, *skip* and *wave1* respectively.

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. Figure 3.4 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 that occur due to similar actions performed by different actors under differing dynamics and/or different actions performed by the same actor.

Yang et al. [92] 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 (please refer to section 2.3.1 for detailed discussion) 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 Figure 3.3(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 classifi-

cation complexity which can be minimized by extracting discriminative information using dimension reduction scheme. Bidirectional 2D-PCA is used instead of other linear/non-linear dimension reduction schemes to retain the correlation amongst adjacent data points as it plays an important role in volumetric data of an action for accurate recognition. Figure 3.5(a) shows better ability of bidirectional 2D-PCA to represent the spatio-temporal information of various action categories performed by actor Daria. Figure 3.5(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 Figure 3.5(a), the separability of different action classes is noticeable whereas components are merged for the feature vectors obtained using PCA (Figure 3.5(b)).



(a) (b)

Figure 3.5: Distinctive features represented among different videos. Spatio-temporal information captured by (a) bidirectional 2D-PCA, (b) PCA.

3.2.3 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, his-

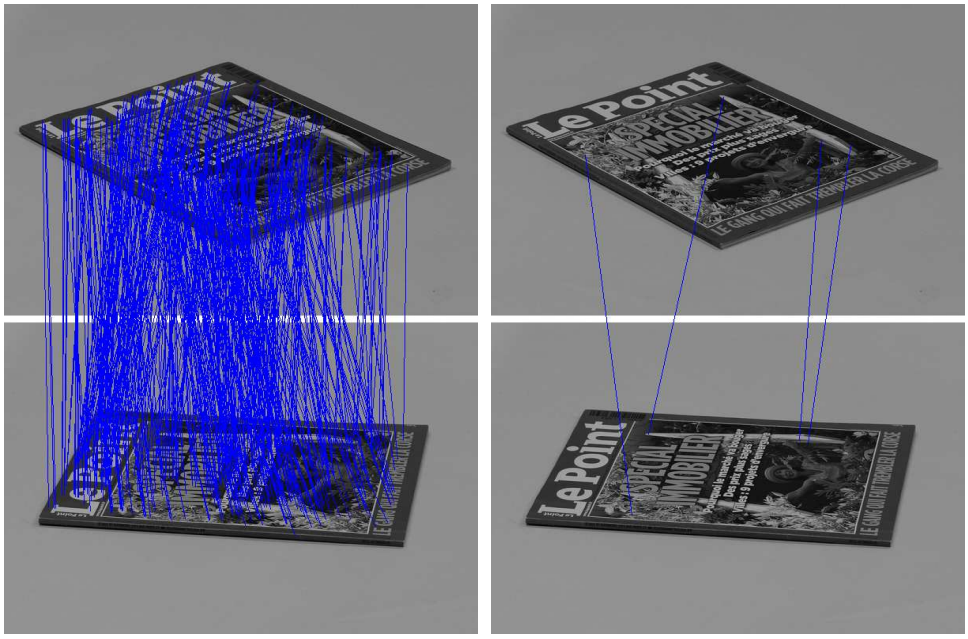


Figure 3.6: Matching of an image pair using ASIFT and SIFT Methods [14]. *Left*: ASIFT matching, *right*: SIFT matching.

togram 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. 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. These methods are not completely invariant to scale changes and affine transformations, however, SIFT performs better than other methods for images with large variations in scale. 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 Figure 3.6). 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.

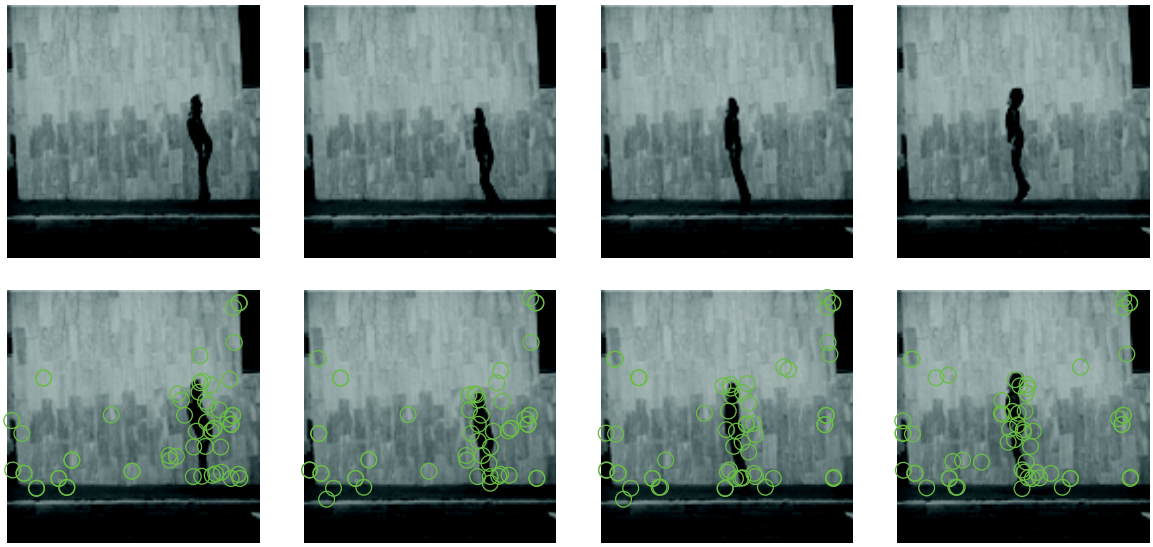


Figure 3.7: Local static features detected in *jump* video of actor *Lena* (green circles represent identified candidate features for matching).

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. Figure 3.6 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.

3.2.4 Pruning of Local Static Feature

We do realize that ASIFT descriptors for patches around interest points may not provide any discriminative information for accurate recognition hence such *local static*

features are eliminated using pruning strategy based on spectral clustering. The pruning operation serves as reduction operation on quantity of *local static* features. A large amount of detected features (bounded in green circles) are shown in Figures 3.7-3.8 for two actions *jump* and *wave2* performed by the actors *Lena* and *Ira* respectively. It is noticeable that the features which are not associated with the moving human body parts do not carry any discriminative information and are justifiably removed.

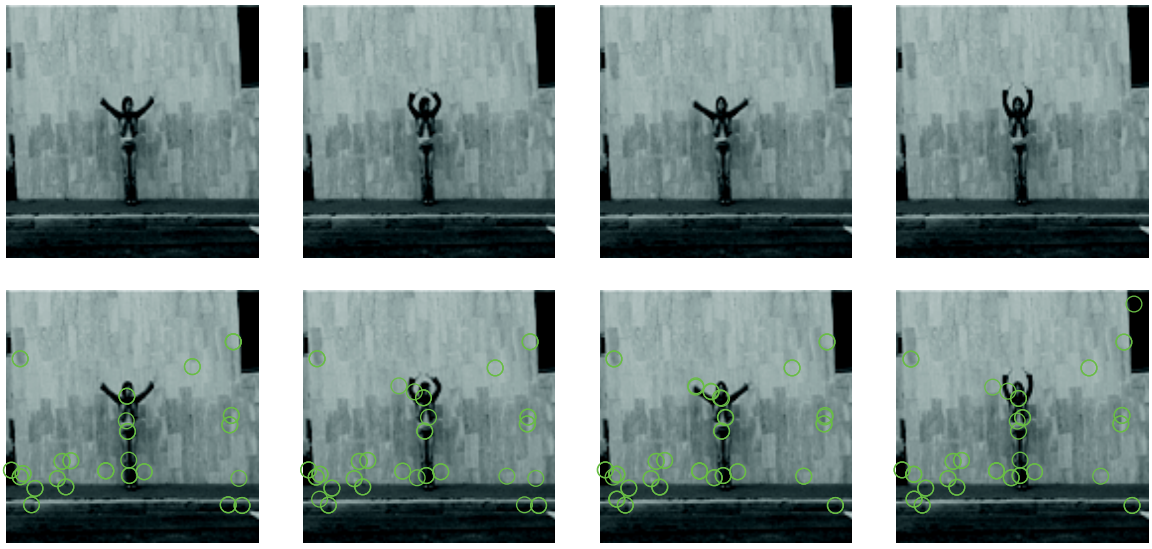


Figure 3.8: Local static features detected in *wave2* video of actor *Ira* (green circles represent identified candidate features for matching).

Spectral Clustering

Clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, engineering, biology to social sciences or psychology. Spectral clustering, belongs to the era of modern clustering schemes, has profusely been used in different areas due its simple implementation and effective performance in finding clusters compared to traditional algorithms such as k-means. Spectral clustering works on the principal of pairwise similarities of the data instances and can be solved using standard linear algebraic techniques

Table 3.1: Spectral clustering algorithm using sparse similarity matrix

Input: Data points $x_1, x_2, ..., x_N$; k number of desired clusters

Output: Cluster Labels for input data instances

1. Construct sparse similarity matrix ζ
2. Compute the graph Laplacian matrix L
3. Computer the first k eigenvectors of L; and construct $V \in R^{N \times k}$ whose columns are the k eigenvectors
4. Compute the normalized matrix U of V using $U_{ij} = \frac{V_{ij}}{\sqrt{\sum_{j=1}^{k} V_{ij}^2}}, i = 1, ..., N, j = 1, ..., k$
5. Use k-means algorithm to cluster N rows into k groups

[38]. Given a set of N data points $x_1, x_2, ..., x_N$; a similarity matrix $\zeta \in \Re^{N \times N}$ is constructed using notion of similarity $\zeta_{ij} \geq 0$ between all pairs of data points x_i and x_j. Similarity graph, $G(V, \sqsupset, W)$ where V and $\sqsupset$ represent sets of vertices and edges, is an efficient representation for data points. Each vertex $v_i \in V$ in the graph represents a data point x_i with its directed connections to other vertices v_j's with edge weights $\zeta_{ij} \in W$ being positive or larger than a predefined threshold. The problem of clustering can be reformulated using the similarity graph, where the objective is to find a partition of the graph such that the edges between vertices of dissimilar clusters carry lower weights and edges within a group carry higher weights. The weighted *adjacency matrix* $W = (\zeta_{ij}), 1 \leq i, j \leq N$ where $\zeta_{ij} = 0$ represents no connection between vertices v_i and v_j. The degree of a vertex $v_i \in V$ is represented as $d_i = \sum_{j=1}^{N} \zeta_{ij}$. The *degree matrix* D is defined as matrix with degrees $d_1, d_2, ..., d_N$ on the diagonal. A subset of vertices and its complement are denoted by $A \subset V$ and $\bar{A}$ respectively; an indicator vector $\mathbb{1}_A = (f_1, f_2, ..., f_N) \in \Re^N$ as a vector with entries $f_i = 1$ if $v_i \in A$ and $f_i = 0$ otherwise. A stream of research is dedicated to graph Laplacian matrices which play a major role in spectral clustering. In the past, dif-

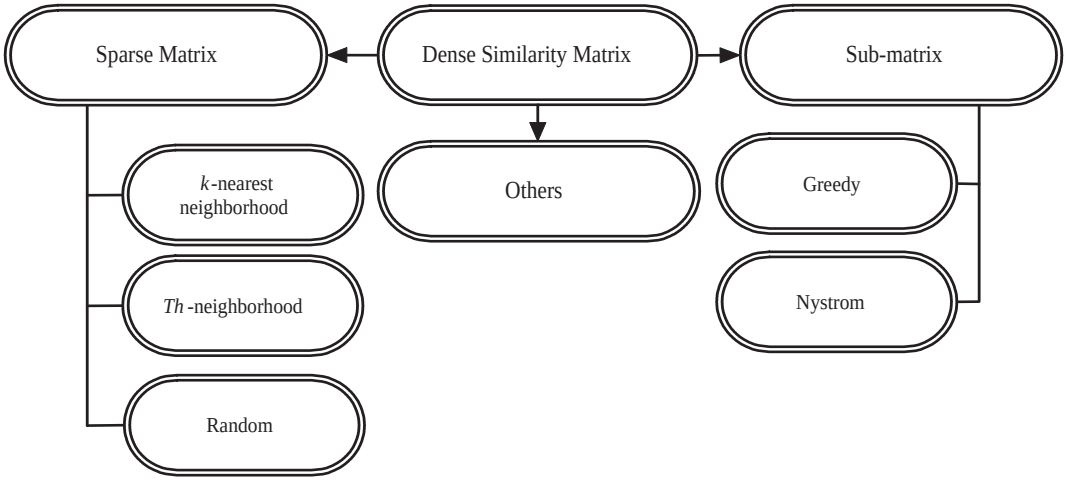


Figure 3.9: Approximation techniques for spectral clustering to minimize storage requirements.

ferent authors have used the term of graph Laplacian for various matrices due to the lack of unique convention. Two widely used representations of graph Laplacian are 1) unnormalized graph Laplacian 2) normalized graph Laplacian. The unnormalized graph Laplacian is defined as $\mathbb{L} = D - W$ while the later type is represented as:

$$\mathbb{L}_{sym} = D^{-\frac{1}{2}}\mathbb{L}D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \tag{3.5}$$

$$\mathbb{L}_{rw} = D^{-1}\mathbb{L} = I - D^{-1}W \tag{3.6}$$

The first matrix $\mathbb{L}_{sym}$ is a symmetric matrix and the second one, $\mathbb{L}_{rw}$, is closely related to a random walk. In the ideal case, where data in one cluster is not related to those in others, nonzero elements of ζ only occur in block diagram form which leads to diagonal graph Laplacian:

$$\mathbb{L} = \begin{bmatrix} \mathbb{L}_1 & & \\ & \ddots & \\ & & \mathbb{L}_k \end{bmatrix}$$

It is obvious that L has k zero-eigenvalues which are also the k smallest ones and their corresponding eigenvectors are defined as $R^{N\times k}$ matrix.

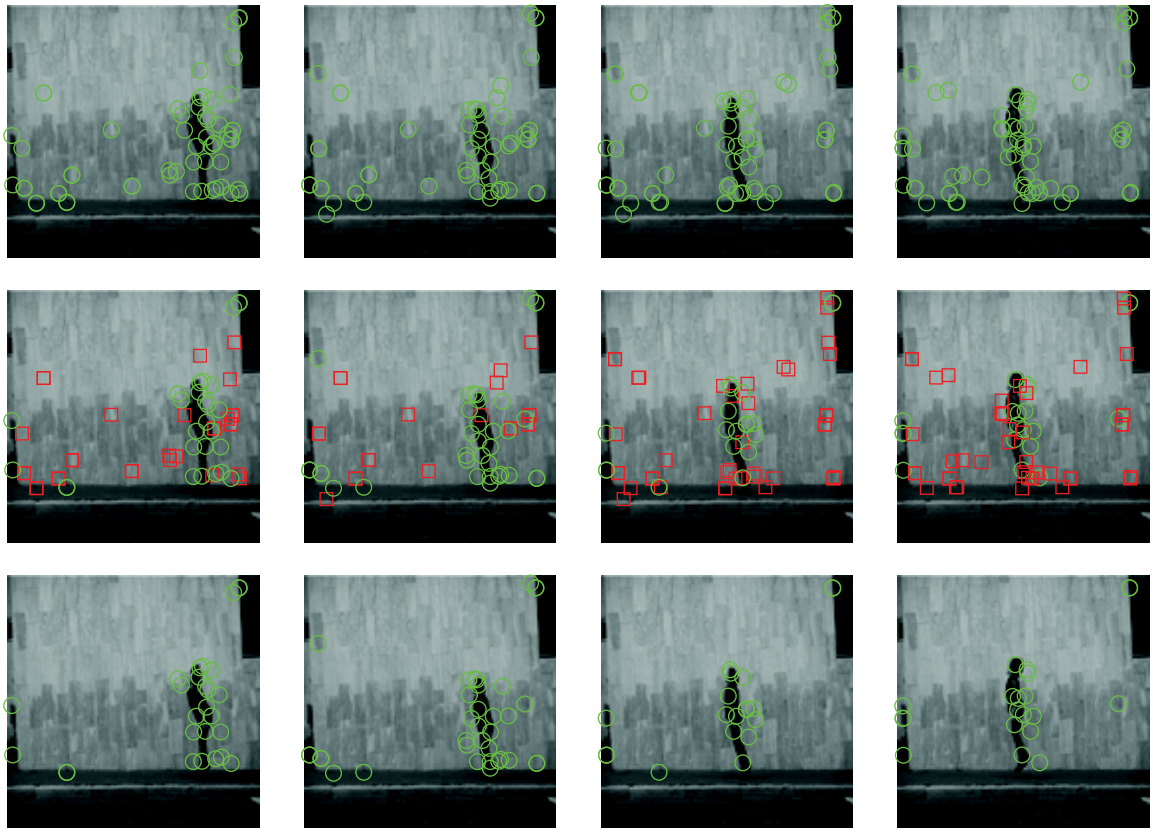


Figure 3.10: Local static features pruned using spectral clustering in *jump* video of actor *Lena*.

Approximation of the Dense Similarity Matrix ζ:

The memory requirement of spectral clustering poses a major bottleneck with elements equal to the square of the number of data points while storing dense similarity matrix ζ. For instance, 8TBytes memory, which is usually unavailable on general-purpose machines, is required to deal with ζ only for 10^6 data instances (assuming double precision storage) [39]. Different approximation techniques have been proposed to avoid storing the dense matrix; Figure 3.9 depicts several existing techniques to sparsify similarity matrix ζ. To generate a sparse similarity matrix ζ, the k-nearest neighbor approach is employed to retain only ζ_{ij} where i (or j) is among the k-nearest neighbors of j (or i). This process reduces the storage complexity of ζ to $O(Nt)$ in-

stead of $O(N^2)$. The computational complexity to construct ζ can also be reduced by $O(n^2 d)$ using KD-tree or Metric trees where d is the dimensionality of data, however, such techniques are not much effective incase of larger values of d. Another possibility to minimize computational burden is to find the neighbors which are close but not the closest. Please be reminded that such approximations may lead to non-symmetric matrices which can be easily converted to symmetric by setting similar values of locations (i, j) and (j, i) if $\zeta_{ij} \neq 0$ or $\zeta_{ji} \neq 0$. Table 3.1 represents main steps involved in spectral clustering using sparse similarity matrix. Readers may refer to [39] to find a detailed discussion on parallel architecture of spectral clustering to avoid the inherent problem of *scalability*.

Feature Ranking via PageRank

This section shows the use of motion cues and PageRank (PR) to extract distinctive local features from the foreground i.e. region of interest. Some videos may have have constantly changing background, thus the local static features detected for such scene areas are not continuously detected throughout the video. The PR approach, successfully used by Google search engine [13], can be used to explore the relatively important and stable features . For an incoming video; a large directed graph of features is generated where a vertex represents a feature and an edge represents a match with another feature. If a feature is consistently matched with many other features , we consider it more significant than others. The idea is similar to consistent feature tracking and PR is a suitable technique to analyze the interaction between the features, by assigning a ranking score to each feature as its relative significance in the feature network.

The discriminative foreground information is not reliably detected between adjacent frames hence we use ASIFT descriptors for a pair of frames (F_t and $F_{t+\tau}, 15 \geq \tau \leq 30$) which are τ (depending upon length of video) time instances apart from each other [13]. Initially, N matched ASIFT feature pairs are estimated, later, a

graph with weight *adjacency matrix* $\mathcal{W}_{N\times N}$ is constructed. Every node of the graph represents a pair of matched features (i, j) and edge weights are computed to measure the geometric consistency of two matches. For instance, if (i, j) and (i', j') are two pairs of candidates then the entry $\zeta_{ij,i'j'} \in \mathcal{W}$ contains the geometric consistency score between two pairs of ASIFT descriptors. Intuitively, correct matches should have a strong correlation with each other while the incorrect ones are random outliers. Now the all the candidate features can be divided into two main groups, i.e. *inliers* and *outliers*, using principal eigenvector of $\mathcal{W}$ which represents spectral clustering solution of candidate matching descriptors [40]. The matching scores of the *inliers* are re-estimated by $\zeta_{ij} = w_1\zeta_{ij}^{geo} + w_2\zeta_{ij}^{app}$ where ζ_{ij}^{geo}, ζ_{ij}^{app} represent geometric and appearance similarity scores respectively whereas w_i represents equal weights assigned to individual similarity score i.e. $w_1 = w_2$. The weights w_i of scores may have different values depending upon their confidence level. The geometric consistency is computed by $\zeta_{ij}^{geo} = \sum_{i',j'\in inliers} \zeta_{ij,i'j'}/vol(inliers)$ which leads to a sparse representation of adjacency matrix W after matching all pairs of frames. Given the constructed large graph G with its vertices and a set of edge weights, the relative importance of the vertices using PR can be computed by treating each vertex as a webpage and all the edge weights associated with the vertex as *votes* cast by the linked neighboring vertices. The features from foreground have higher number of consistent matches throughout the video sequence leading to higher *votes* compared with features detected in changing background. The PR values of individual features are represented by $Pr_{1\times N}$ which is computed by

$$Pr = \alpha \times Pr \times W + (\alpha \times Pr \times b + 1 - \alpha) \times v \tag{3.7}$$

where α is the scaling factor, b is an indicator vector identifying the vectors with zero-out degree, and $v_{N\times 1}$ is the uniform probability distribution over the vertices. The initial PR value for each vertex is $1/n$. A PR vector Pr_t is computed for each frame F_t; only top μ features are selected based on the rank of Pr values. The qualitative

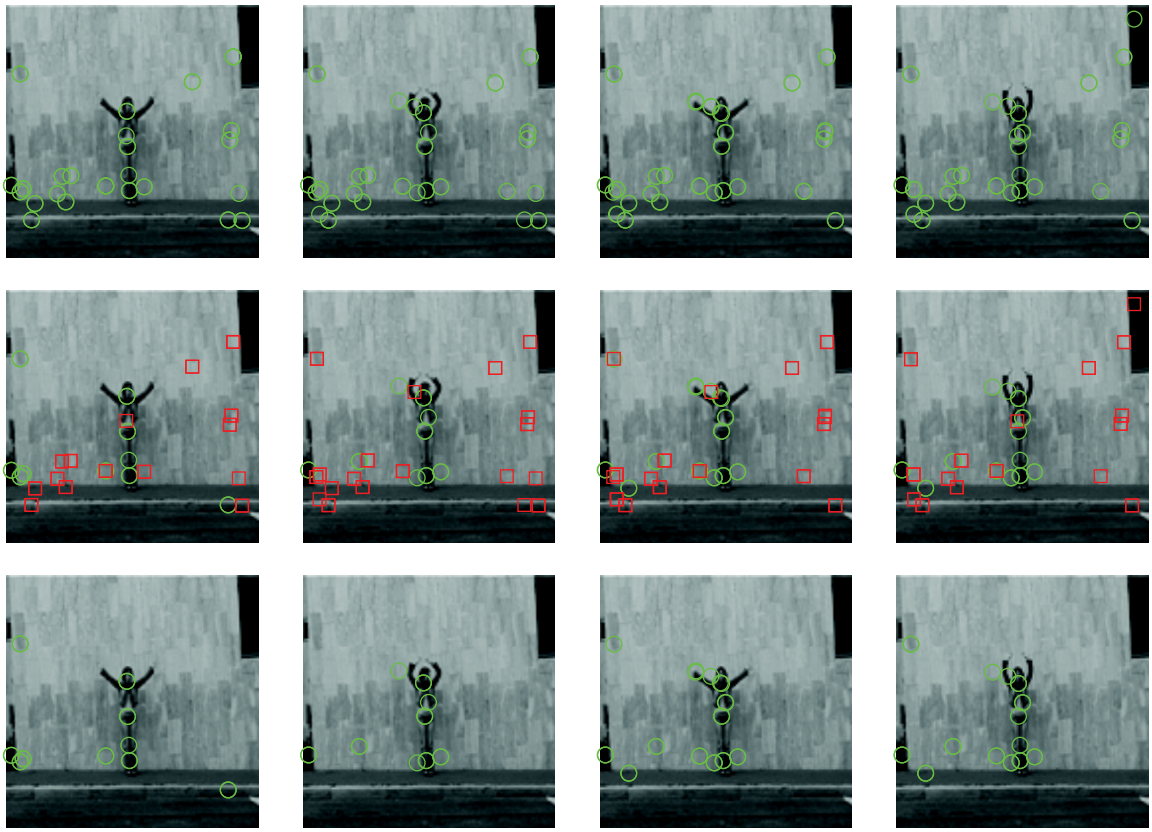


Figure 3.11: Local static features pruned using spectral clustering in *wave2* video of actor *Ira*.

performance of pruning strategy is shown in Figures 3.10-3.11 with eliminated *local static* features bounded in red squares whereas the retained cluster of highly discriminative feature sets are bounded in green circles to be utilized in further recognition.

3.3 Results and Discussion

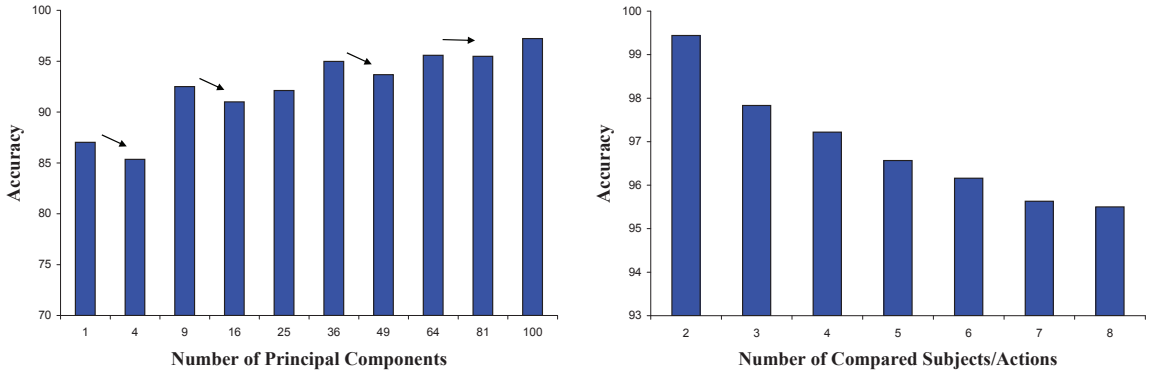
To test the performance of our proposed method, publicly available datasets, Weizmann [6] and KTH [11], are used in our experiments. It is pointed out that the results presented cannot be considered as direct comparison against other recognition schemes because of all kinds of variations of the experimental setups and assump-

tions about *a priori* knowledge of the video/action being investigated. However, 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. We executed all of our experiments in MatLab environment on an Intel Core 2 Duo processor of 1.80GHz speed and 2GB RAM.

Important advantages offered by our proposed scheme include no requirement of video alignment and the amount of feature vectors proportional to the number of frames and the level of multiresolution decomposition. We apply our classification scheme using *spatio-temporal* features 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 Figure 3.12(a), accuracy is not constantly increasing by raising dimensionality of feature vectors; especially classification precision dwindls or remains constant at arrow locations as dimensionality increases. 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. Figure 3.12(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



(a) (b)

Figure 3.12: Accuracy analysis (a) using *spatio-temporal* features (of varying size) only (b) varying number of compared subjects/actions using hybrid features.

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 Figure 3.12(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 Figure 3.12(b) correspond to the average classification accuracy achieved for Λ number of subject test videos. The classification accuracy presented in Figure 3.12(b) is obtained using hybrid features and it is apparent that the trend line of achieved ac-

curacy 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, apparently the same activity in between repetitions of an actions is observed in a small number of adjacent frames.

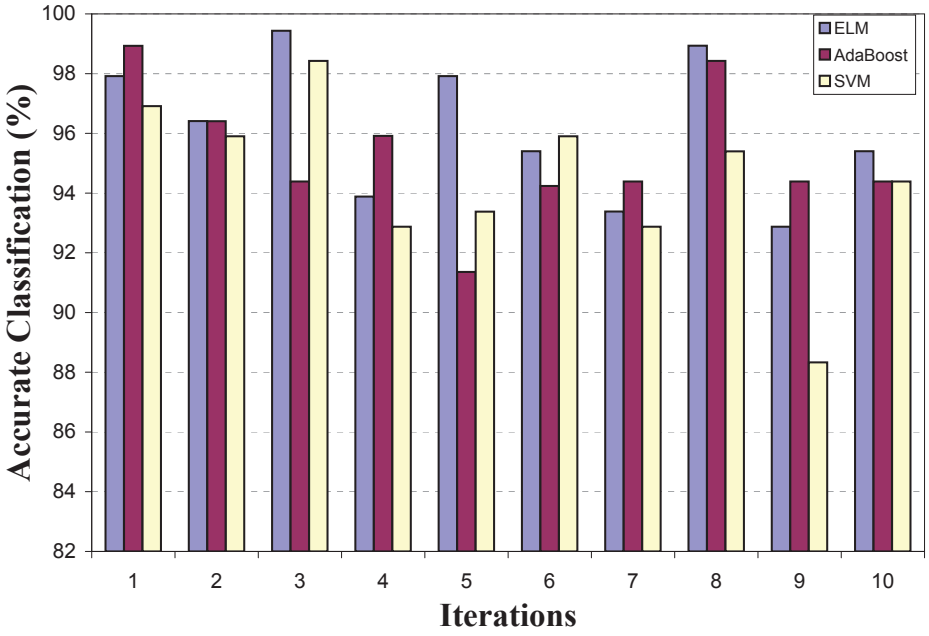
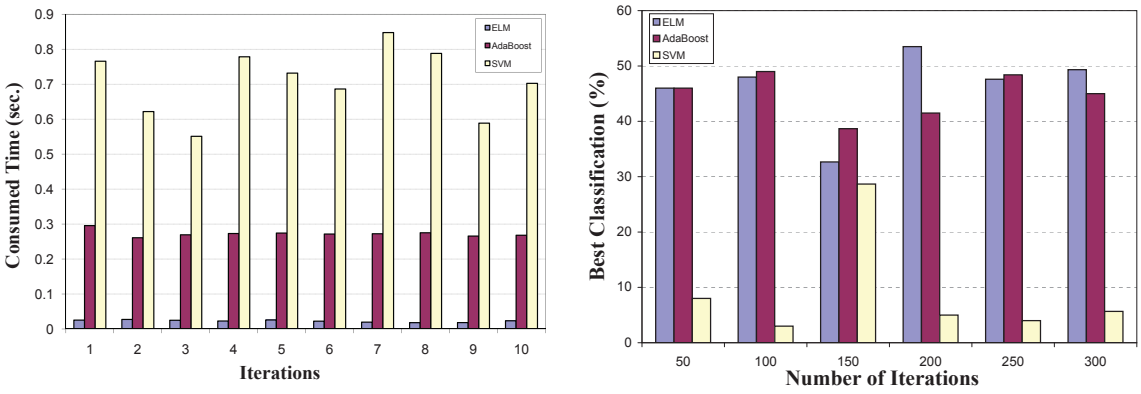


Figure 3.13: Accuracy analysis of different classifiers using hybrid features extracted from Weizmann dataset.

3.3.1 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



(a) (b)

Figure 3.14: Performance analysis of different classifiers using Weizmann dataset (a) computational complexity analysis (b) best classifications achieved for varying iterations.

(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 multiclass problems. Accuracy of classification is illustrated in Figure 3.13, 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 Figure 3.14(b) in terms

of collective percentage of best classifications achieved by individual classifiers.

In terms of computational complexity, as shown in Figure 3.14(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

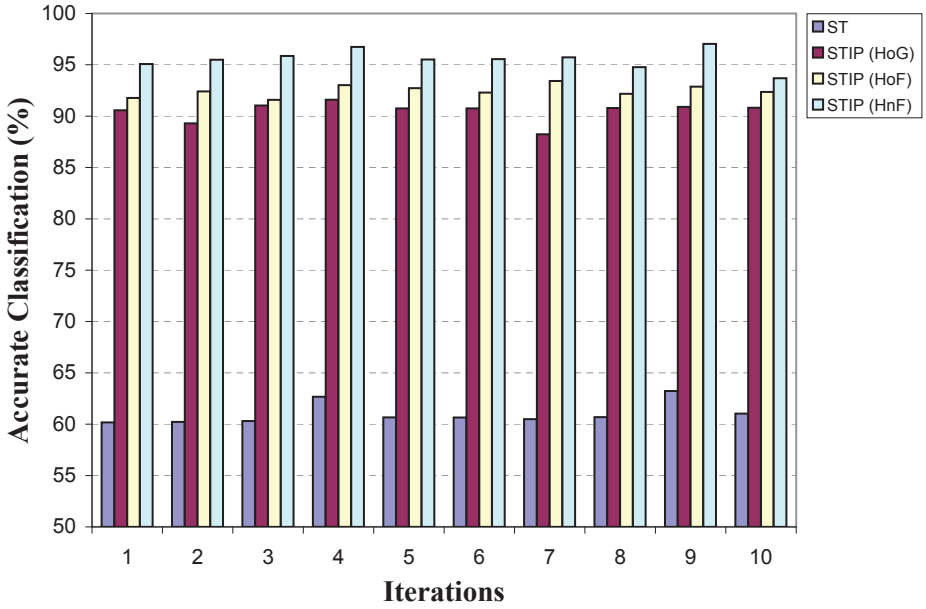


Figure 3.15: Performance analysis of ELM using various *spatio-temporal* features for Weizmann dataset.

datasets ELM is employed to four types of spatio-temporal features namely ST, STIP (HoG), STIP(HoF) and STIP (HnF). Figure 3.15 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%

Table 3.2: 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

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.

3.3.2 Performance Analysis of Proposed Framework

Table 3.2 shows 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 classificatin problem

Table 3.3: 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

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 3.3) it is noticeable that classification uncertainty is present among two action subgroups. Actions involving hand movements such as *box, wave* and *clap* belong to the first group whereas the second class of actions consists of legs/feet motions (*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*. The classification precision of 94.83% is achieved for KTH datasets, which demonstrates favorable results for proposed action recognition scheme. Figures 3.16 and 3.17 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. In addition to improved accuracy, our proposed scheme does not require any *a priori* information such as foreground masks for segmentation. On all video sets, our approach renders improved and/or comparable recognition accuracy against existing schemes. In Figure 3.16, 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 Figure 3.17.

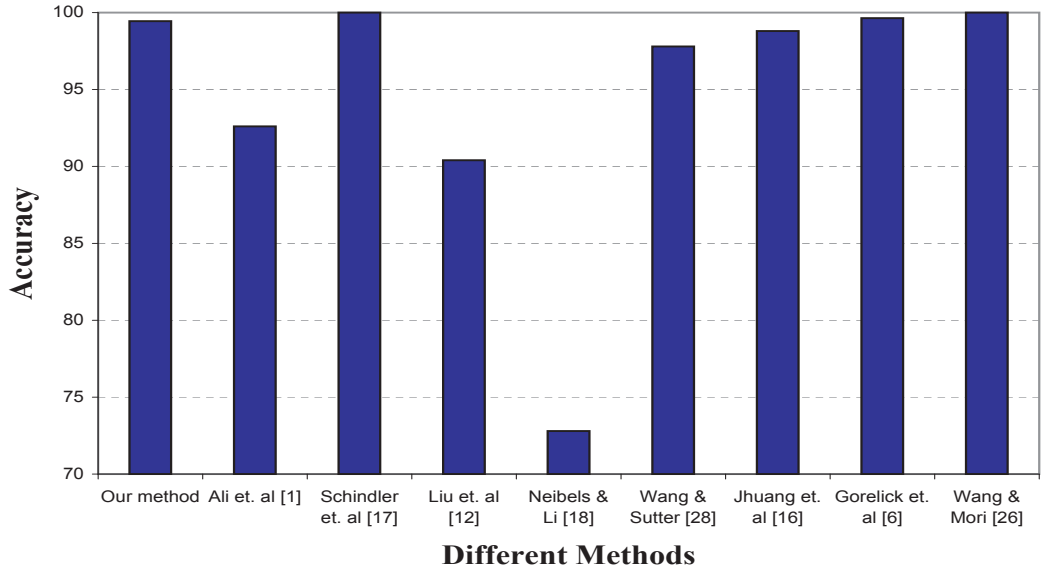


Figure 3.16: Performance comparison for various methods using Weizmann datasets.

3.3.3 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 daily life; we test our algorithm on 10 different styles of walking from the same view point. Figure 3.18 presents sample video frames from the dataset under-reference where clutter background, partially obscured human body due to skirt, pole, and box complicate the classification task. Another kind of activity videos, termed as *robust view*, are 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. Figure 3.19(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

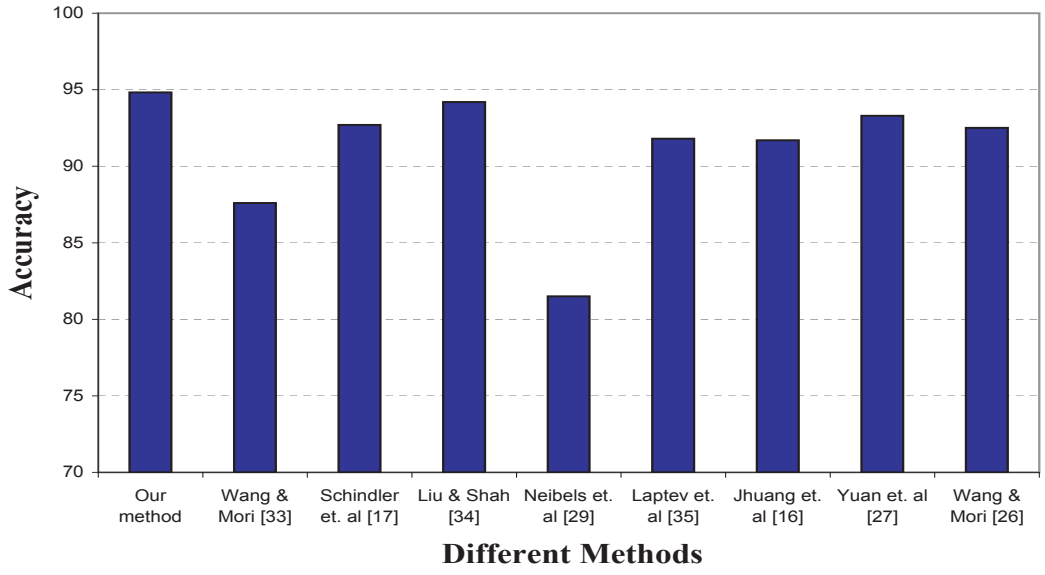


Figure 3.17: Performance comparison for various methods using KTH datasets.

changes, however, it exhibits robust behavior in presence of partially occluded objects, scale changes and non-rigid transformations. Figure 3.19(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.

Figure 3.18: Sample images from Weizmann robustness datasets. Left to right: *with dog, with bag, knees up, pole, in skirt* and *no feet* action videos.

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

(a)

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

(b)

Figure 3.19: Robustness evaluation of our proposed method using Weizmann robustness datasets (a) details of dataset (b) recognition comparison for different techniques i.e. [6], [28] and our method (top to bottom).

3.4 Summary

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.

Chapter 4

Recognition through Recursive Training

Action recognition within a video sequence is a dynamic area of interest for researchers due to its imperative applications in both military and non-military problems. Most published methods in action recognition rely on impractical assumptions such as the processing of an entire video or require a large look-ahead of frames to label an incoming video. Based on an extensive literature survey incremental learning is an often overlooked obstruction in the implementation of recognition frameworks which employ real-time yet powerful classifiers. Schindler and Gool's work [17] utilizes *snippets* of length $1 - 10$ frames, this important breakthrough does not rely on impractical assumptions. These *snippets* are used to extract shape and optical flow information. Schindler and Gool's reported results are based on a bank of linear classifiers and multiple types of features, which are comparable to the methods using entire videos.

4.1 Introduction and Challenges

This chapter's objective is to systematically find a recognition method which adaptively utilizes the least possible information accumulated over the past few video frames. Such findings may significantly help to realize action recognition schemes for real life problems with lower computational complexity. Existing frameworks are broadly criticized for using higher amount of information than required in recogni-

tion. Based on recognition capabilities of the human visual system, one can argue that we do not require an entire video to identify an ongoing event. Additionally, the use of a sub-sequence instead of an entire video can help to identify actions which are combinations of more than one actions, e.g., *entering into a building* may entail *walking, opening a door* and then *walking* again. Most of the existing approaches fail in such conditions since the features from the entire video may misrepresent the action as being *walking* or *opening a door* instead of *entering into a building*. To refer to a sub-sequence of a video the term *snippets* [17] is used for simplicity reasons and is congruent with existing nomenclature.

The action recognition frameworks with online sequential learning are far from reality due to the fact that the training phase in state-of-the-art schemes are assumed to be offline. Such batch mode training holds back the application of subsisting frameworks to the problem where events are evolutionary. For example, an exercising person on a treadmill may *walk*, *jog* and finally start *running*. All the actions he gradually performs are associable because of some common characteristics, hence the learnt model needs to be updated as the time passes. Existing techniques must re-learn the entire model to differentiate amongst evolving actions which are unseen in earlier training. Such retraining is a time consuming process which is comprised of numerous iterations through the training data. In traditional learning paradigms, learning a model may take from several minutes to hours and the learning parameters (i.e. learning rate, number of ensembles, stopping criteria and other predefined learning constraints) must be carefully chosen to ensure convergence. Additionally, whenever a fresh training data is received, batch learning uses the past data together with the new data and performs a retraining. Intuitively, online sequential learning algorithms present a preferred solution for a generic action recognition scheme.

This chapter's main emphasis is to find the solutions of three overlooked but important problems *1)* How to efficiently represent global and partial information along with spatial layout? *2)* How to minimize the training time that poses a major bot-

tleneck in recognition tasks? *3)* How to extend the idea of online sequential learning to action recognition frameworks, which may help to avoid the relearning of existing information. It should be noted that human or OI are terms used interchangeably to refer to a target object.

4.2 Proposed Recognition Framework

The motivation behind this chapter is to design a framework that utilizes the minimum number of video frames for action recognition. The proposed recognition scheme does not require an entire input video, instead, a small collection of frames can be used to identify an action. Recent work by Schindler and Gool [17] presents an excellent introduction to the subject with promising results achieved by explicitly extracting the shape and the optical flow information between consecutive frames. This chapter borrows similar motivation and extends this idea for incremental learning. The extraction of features exploiting only a small collection of frames, called *snippes* by Schindler, eliminates the need for a large look-in-advance to recognize an action. This line of exploration has many advantages such as online action recognition with lower computational burden and identification of events which may have occurred in a small portion of the video. The proposed framework offers two important contributions 1) action recognition using a small number of frames through tracking based on shape and appearance variations 2) incremental learning of the proposed framework with generalized performance and training analytically performed at extremely fast speed. In general, a smaller video sequence consists of a lower content of information for decision making. However, the useless information of background and static objects in the scene presents additional challenges. As mentioned earlier the motivation behind this work is inferred from [17], however, the proposed framework is significantly different from Schindler's work in the following ways: 1) unlike [17], there is no need to trim videos to account for uneven lengths 2) the proposed scheme

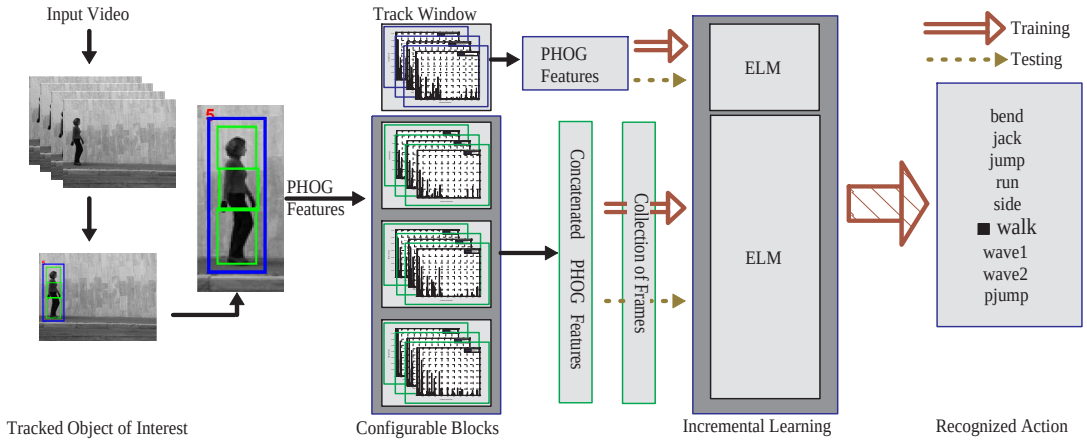


Figure 4.1: Overview of proposed recognition system learned incrementally using PHOG features extracted from adaptive blocks to approximate contour of a moving object.

is based on incremental learning which offers better flexibility adopting new classes of actions without the obligation of rerunning the entire training phase 3) the extracted features are computed from annotated body parts undergoing rapid variations 4) the dimensionality of feature vectors are less than half of what is required by [17] which minimizes the computational load. The proposed framework is able to classify actions with speeds as fast as 4-5 frames per second with implementation in the Matlab environment. This performance can be further improved by changing the execution platform. In addition, the simpler structure of ELM contributes to improved recognition, minimum false alarms and higher training accuracies. The proposed method is essentially operable for heightened situations such as *snippets* of lengths ranging from one frame to an entire video.

4.2.1 Input Data

The input of this algorithm is the single video frame with the contour initialization as the lone pre-processing step. Additional video trimming, stabilization or segmentation based upon readily available masks are not performed. The initialization of

contour is the only primitive attention mechanism needed to start classification. It is worth realizing that the proposed framework can easily be extended to recognize multiple actions in contrast to existing schemes that handle a single dominant action [5],[16],[17]. There is a mentionable difference between our initialization setup and other state-of-the art. Our method relies on foreground segmentation of a small window that encapsulates the person of interest based on shape and appearance information, whereas, Efros et al. [5] localize windows assuming uniform background and thus extract relative articulations of the human body. The scheme in [16] uses a person-centered frames with additional motion information being obtained through the image global coordinates while observing the inverse flow of the background within a stabilized window. In comparison to silhouette based schemes [6],[28],[46], bounding box approaches offer further generality since a reliable silhouette extraction requires static background. On the contrary, bounding boxes are naturally obtained using person detectors based on sliding windows [109].

Based on the length of *snippets*, recognition accuracy significantly varies, intuitively, higher number of frames tender more information that leads to reliable recognition. Following initialized contour, the articulated target object is tracked in subsequent frames and represented with a large box called as tracking window and rectangular blocks that reside inside the tracking window to approximate human contour. As a point of fact, the track window and the rectangular blocks correspond to bounding boxes for overall human body and the various body parts, respectively. Finally, Pyramid of Histogram of Oriented Gradients (PHOG) features for rectangular blocks are concatenated to form a single higher dimensional feature vector per video frame followed by learning and testing using recursive ELM. The main steps involved in our proposed framework are depicted in Figure 4.1. The tracking and representation of a moving human using PHOG is represented in the following section. Please be reminded that the dimension of PHOG feature vectors computed using track window and configurable blocks are of different sizes which are trimmed to the maximum

dimension of 100 due to two reasons 1) to use a single ELM for testing, provided that available computational resources are limited 2) we do not achieve mentionable improvement in accuracy for higher dimensional feature vectors.

4.2.2 Adaptive Representation of Human Body as PHOG Features

A visual tracking methodology deals with a consistent identification of a feature point or an OI in an input sequence irrespective of variations in shape and appearance. The implementation of a precise and robust visual tracker is a challenging task whose complication increases further when an OI undergoes large and rapid shape and appearance variations. Generally, the change in appearance is mainly due to change in shape while the foreground intensity distribution roughly stays stationary. This assumption of weak appearance constancy can be exploited for accurate tracking of a movement in a video sequence. A simplest way to track an OI based on appearance utilizes intensity histogram and the concept of integral images to spot rectangular shapes has also been successfully used in the past [77],[122]. However, intensity histogram can not be applied to track shapes varying in an irregular manner. The use of intensity histogram for an entire image may also deteriorate the performance of a tracker if the image size is too big and/or the size OI is too small. Usually, it is feasible to scan an entire image to locate the position of an OI, even with irregular shape, and enclose it in a box for intensity histogram computation. Subsequently, such representation still consists of background pixels resulting in corrupted and non-distinctive feature vectors [72].

Shape and Appearance based Tracking

The spatial layout and geometric information play a pivotal role in tracking an OI going through rapid shape changes. The unembellished histogram becomes insuffi-

cient to recognize such OIs and yields unstable tracking. The concept of *spatiogram* provides partial solutions to above concerns at the cost of higher computational complexity that leads to tracking performance unsuitable for real time applications. Our proposed action recognition is based on adaptive tracking utilizing changing shape and appearance of the entire human body and its parts. The tracking module in our recognition scheme attempts to model gradually changing shape of an OI (foreground is the alternative term used in this chapter), therefore position and size of small rectangular blocks within a track window are adaptively changed.

The tracking module in our recognition framework consists of global scanning, update of intensity histogram based on entire human body and its annotated parts to closely trail its contours [72]. The appearance is represented by the histogram that can be easily determined to locate an OI while scanning an entire image. Shape update is carried out by adjustments of a few blocks within a track window. The use of such small blocks is helpful to approximate the uneven shapes through integral histograms. The blocks inside the track window cover the majority portion of an OI with minimal overlap. The size of tracking window is typically small enough to perform a fast segmentation for contour extraction of the target object, i.e. moving human in our case (see Figure 4.2). Finally, the target shape is updated by tuning the locations and sizes of these blocks so that they can provide maximal coverage of the OI. The arrangements and associated weights of the blocks are adaptive. The shape of an OI is estimated by the structure whereas appearance is represented using intensity distribution and associated weights of the blocks. Exact representation of shape and appearance is not the goal of the tracking module; rather our focus is to approximate the varying shape and appearance as a result of human movement. The weak constancy assumption for appearance helps to reliably find a track with least computational load. The difficult part of intensity histogram based tracking is the approximation of foreground object histogram under significant shape variations. The tracking part consists of three sequential steps 1) detection 2) refinement and

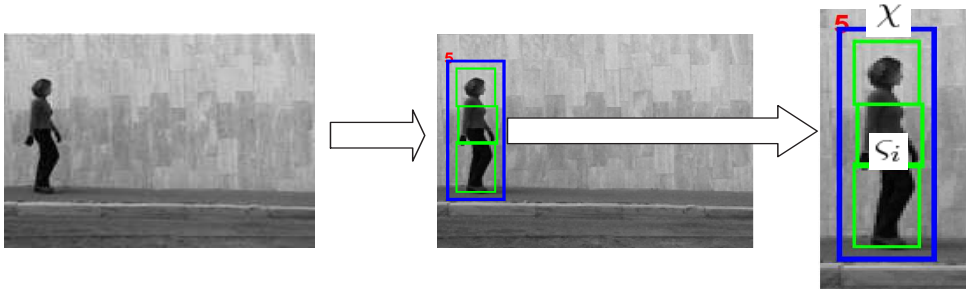


Figure 4.2: An articulated OI (left), given the contour a track window χ (in blue) and adjusted blocks ς_i (in green) to approximate shape.

3) update. As a starting point, the tracker is initialized with the contour of the target which automatically determines the tracking window χ and 3 rectangular blocks ς_i along with their associated weights w_i. The number of blocks represent a trade-off between computational resources and required accuracy. It is noticeable that all the reconfigurable blocks ς_i are located inside track window χ that corresponds to a bounding box for the contour of a moving OI. The foreground intensity histogram $\mho_0^{\mathrm{F}}$ for the initial frame and the positions of the blocks ς_i are maintained throughout the processing with minor overlap to account for rapid shape changes due to movements, as shown in Figure 4.3. For each time instance t, we maintain the followings a) a template window χ_t with a block configuration b) a foreground histogram $\mho_t^{\mathrm{F}}$ computed using local histograms $\mho_t^{\varsigma_i^{\mathrm{F}}}$ of the blocks and their associated weights c) a background histogram $\mho_t^{\mathrm{B}}$. For each frame, the entire image is initially scanned to locate a window $\tilde{\chi}$ that encloses the target object. The windows χ_t and $\tilde{\chi}$ should satisfy the maximum similarity criterion $\tilde{\chi} = max\, \mathbb{S}(\acute{\chi}, \chi)$ where $\acute{\chi}$ ranges over all the scanned windows. The similarity measure $\mathbb{S}$ is computed by overlaying block configuration template window χ to each of scanned window $\acute{\chi}$ and accordingly evaluating local histograms of the transferred blocks. The local foreground histogram $\mho_t^{\varsigma_i^{\mathrm{F}}}$ for the block ς_i is the intersection of the raw histogram $\mho_t^{\varsigma_i}$ with the initial initial histogram $\mho_0^{\varsigma_i^{\mathrm{F}}}$ of the corresponding block. The similarity measure is defined as the weighted

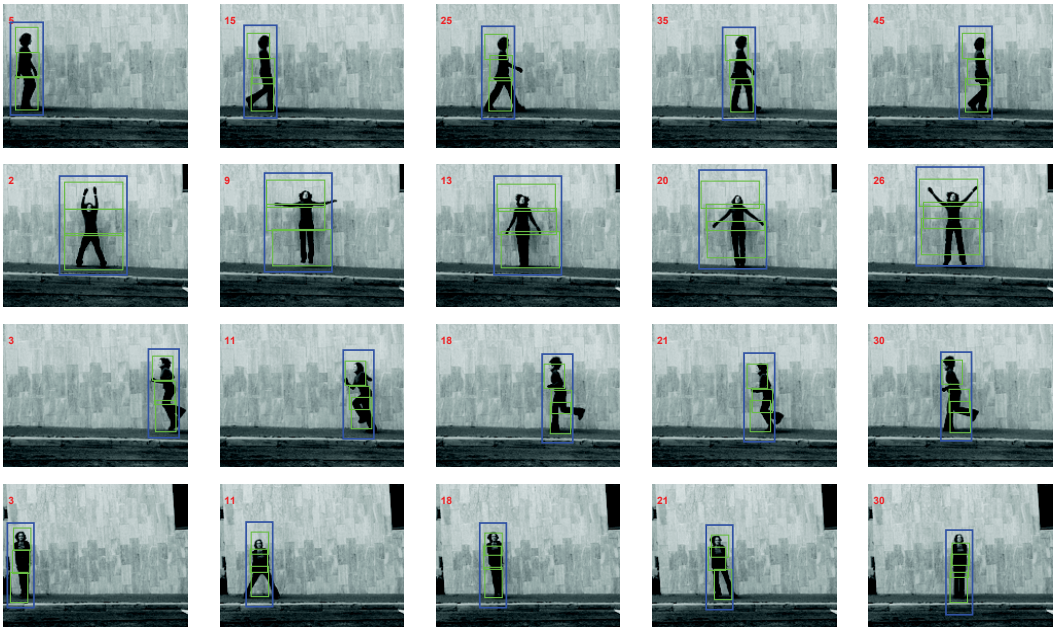


Figure 4.3: Tracking results using action videos of *walk, jack, skip* and *side* (top to bottom) performed by actor *Lena*.

sum of the histogram similarities.

$$\mathbb{S}(\acute{\chi}, \chi) = \sum_{i=1}^{3} w_i \mathrm{D}(\mho_t^{\varsigma_i^{\mathrm{F}}}, \mho_0^{\varsigma_i^{\mathrm{F}}}), \;\; \mho_t^{\varsigma_i^{\mathrm{F}}} = min(\mho_t^{\varsigma_i}(b), \mho_0^{\varsigma_i^{\mathrm{F}}}(b))$$

where b and D correspond to the indexes of histogram bins and Bhattacharyya distance, respectively. The above weighted sum limits the contributions of blocks ς_i with higher number of background pixels since such mechanism is required against cluttered background and noise. Additionally, such configuration of blocks ensembles a degree of spatial information and shape of the foreground object unlike other schemes that exploit only one histogram for an identical matching problem.

The next step after locating a target window is the extraction of approximate body contour using graph cut segmentation on $\tilde{\chi}$. The traditional cost minimization function for such segmentation is based on appearance as well as shape. However, individual term for shape without dynamic information in cost minimization is a bottleneck towards improved recognition due to extreme variations in shape. Our

solution aims to minimize the use of appearance term by incorporating shape information through foreground block densities [53],[72],[123]. Since graph cut is applied in a small window $\tilde{\chi}$ the computational load is minimum. The extraction of contour leads to adjustments in positions of blocks ς_i within tracking window $\tilde{\chi}$. The obvious goal is to configure blocks in such a way that a maximal coverage of segmented foreground object is achieved. Simultaneously, these blocks are adjusted in a controlled elastic manner to account for large articulated motions. A greedy strategy, exploiting size based ordering, is applied by moving blocks to attain maximum coverage of the segmented OI. Since the foreground definition is now known, the histogram inside blocks can be determined with few additions using integral histogram. Finally, to maintain a stationary foreground density of the initial frame, $\mho_0^{\mathrm{F}}$, following weighted sum relation is used:

$$\mho_0^{\mathrm{F}} = \sum_{i=1}^{3} \mathrm{G}_i \mho_t^{\varsigma_i^{\mathrm{F}}}, \; w_i = \mathrm{G}_i \varpi_i,$$

where $\mathrm{G}_i \geq 0$ and ϖ_i represents the percentage of foreground pixels in ς_i.

Shape Encoding with Spatial Pyramid Kernels - PHOG Features

This part explores the effects of spatial layout, missing in visual vocabularies, of descriptors in action recognition process. As per the knowledge gathered from our surrounding, some of the objects are geometrically quite constrained whilst others have greater variations such as human beings. Despite shape variations, since the local shape is adaptively extracted using shape and appearance based tracking that allows us to efficiently symbolize these extracted image regions taking full advantage of their spatial layout. The PHOG features present a relatively fresh idea based on spatial pyramid kernel that uniquely represents *local* image shapes and their spatial layouts. The main goal is to transform image regions into distinctive descriptors which can be used at later stage for classification. PHOG descriptors are mainly inspired by two different sources a) the Histogram of Oriented Gradients (HOG) features [109] b)

the image pyramid representation of Lazebnik [124] which argues that a strong match goes beyond a *bag-of-words* and involves the spatial correspondence. The local shapes are captured by the distribution of oriented edges within a region, and spatial layout by tiling the image into regions at multiple resolutions (Figure 4.4). The descriptor consists of a HOG over each image subregion at each resolution level. The local shape of an OI is encoded by the histogram of edge orientations quantized into specified bins where the contribution of each edge is weighted according to its magnitude. Each bin in the histogram represents the number of edge with orientation residing within a certain angular range. For spatial layout, at individual pyramid level we compute

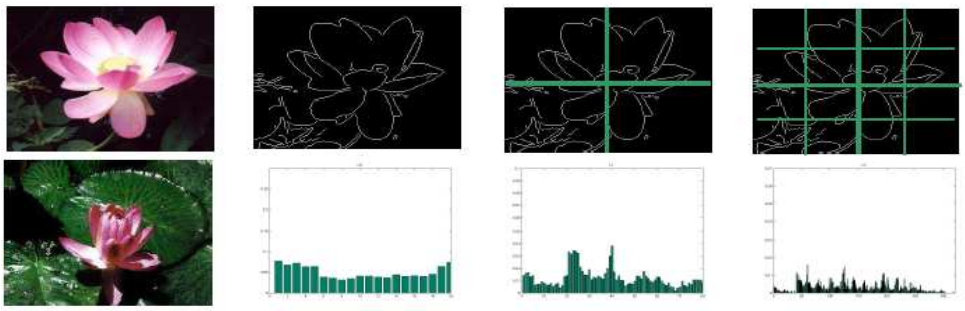


Figure 4.4: Shape spatial pyramid representation [103]. *Top:* an image with its grids for levels 0,1 and 2 (left to right). *Bottom:* an image with its histogram representations for corresponding levels.

HOG feature for each grid cell and concatenate them to form an extended PHOG feature vector. The PHOG is normalized to sum to unity to ensure that images with more edges do not get higher weights compared to other plain images. In principal, the computation of PHOG is justified because of 1) insensitivity to small rotations 2) compact vector representation 3) the ability to cope with varying degrees of spatial correspondences. Bosch et al. [103] have presented a generalized technique to embed shape as well as appearance information based on weighted strategy using kernels. However, extracting appearance information using SIFT [10] at points of regular grid with image patches of varying radii is computationally expensive and requires larger

memory, especially, for multi-channel color images, i.e., HSV. In our experiments we do not require appearance information while computing PHOG features, firstly, the appearance has already been matched in tracking part to locate annotated human body parts. Secondly, it is a time consuming process that deteriorates real-time performance of our proposed recognition framework. Provided the tracking results

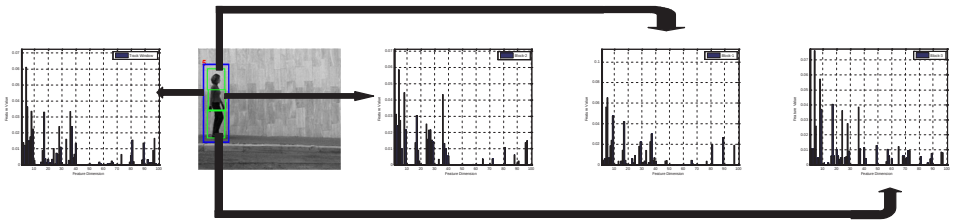


Figure 4.5: A tracked OI with its computed PHOG features for the track window χ (in blue) and blocks ς_i (in green) to approximate the shape.

shown in Figure 4.2, the computed PHOG features for track window and blocks are presented in Figure 4.5. It is obvious that a considerable portion of track window area consists of background information which does not provide any information for reliable recognition. The enclosed background portion largely varies based upon the type of action being observed. Please refer to Figure 4.3 where track window χ for action *jack* chiefly bounds background area compared to what is covered for the foreground object whereas this information proportion varies opposite way for other actions performed by the same actor. This situation demands us to either discard PHOG features for track window completely or give it less weights for final classification due to lower confidence. In view of the fact that occasionally large portion in a track window comprises of only foreground object which ultimately leads to valuable PHOG features; the contribution of these feature vectors is adjusted to lower level in *snippet* classification. It requires special mention that foreground object are not always completely bounded or covered by the track window χ since the tracking strategy tries to provide maximal coverage of the OI, however it can

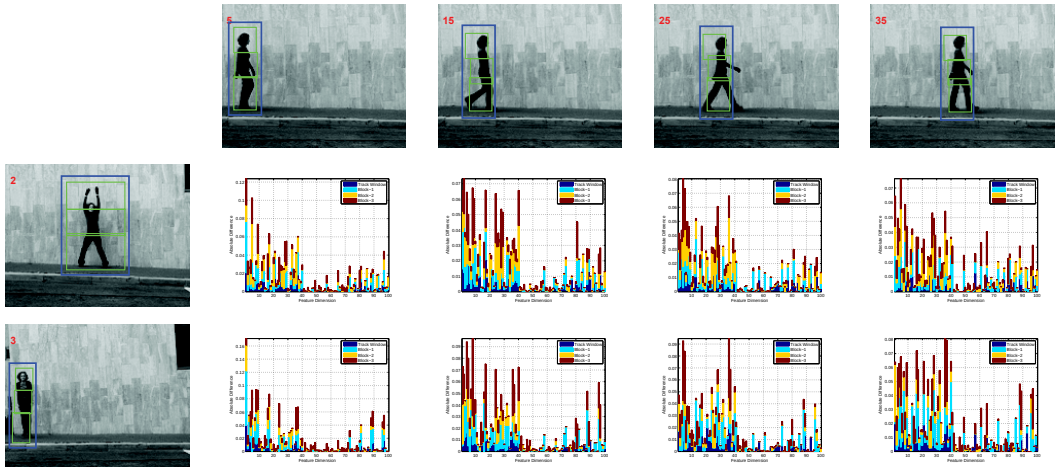


Figure 4.6: Difference of PHOG features for action videos between *walk-jack* and *walk-side* (top to bottom) performed by actor *Lena*.

sporadically fail to spot small body portions because of dire contour initialization and speed of an action or huge shape variations (see Figure 4.3). Although, omitted body parts, such as foot and hands etc., do not adversely degrade recognition since spatial layout, blocks overlap, shape and appearance information have already been integrated for action recognition. A perceptive imagination reveals the fact that the match amongst PHOG features for dissimilar objects and/or actions eventually leads to an unreliable recognition which is a likely phenomenon for track window χ in situations where a large portion consists of scene background (action *jack* for Figure 4.3). A similar behavior is observed in Figure 4.6, where the absolute difference for different actions is considerably lower for features from track windows χ compared against rectangular blocks ς_i.

4.2.3 Recursively Trained ELM

The back propagation (BP) algorithms and its variants have been vital schemes for training of FNNs. It is to be noted that BP is basically a batch learning algorithm whose main variant, stochastic gradient descent BP (SGBP) represents a sequential

learning approach for training. The network parameters of SGBP are tuned at each iteration on the basis of first-order information of instantaneous value of cost function using the training pattern [70]-[71]. To overcome the slow convergence and shorten the convergence time, algorithms based on second order information for network parameter learning have been proposed. However, such schemes require additional time to process individual training patterns that can degrade the overall performance of online learning whereas network size of SGBP needs to be predefined and fixed, in advance.

The learning of ELM (Section 2.2) is simply equal to finding a least-square solution of $|\Upsilon(\hat{w_1},\ldots,\hat{w_L},\hat{b_1},\ldots,\hat{b_L})\hat{\beta}-\Gamma\| = \min_{w_i,b_i,\beta}\|\Upsilon(w_1,\ldots,w_L,b_1,\ldots,b_L)\beta-\Gamma\|$. The smallest training error is achieved by using above model since it represents a least-square explanation of the linear system of $\Upsilon\beta=\Gamma$ as $\left|\left|\Upsilon\hat{\beta}-\Gamma\right|\right| = ||\Upsilon\Upsilon^{*}\Gamma-\Gamma\,|| = \min_{\beta}||\Upsilon\beta-\Gamma||$ where Υ^{*} represents *moore-penrose* generalized inverse of hidden layer output matrix Υ. The above solution assumes that all N distinct training observations are available which is a typical situation for batch learning. However, in real-life applications such data may arrive in sets varying anywhere from 1 to N [71]. Hence, the batch ELM algorithm needs to be modified to fit the requirements of online sequential learning. Under the condition of $rank(\Upsilon)=L$

$$\Upsilon\beta=\Gamma,\quad \beta=\Upsilon^{*}\Gamma,\quad \Upsilon^{*}=[\Upsilon^{T}\Upsilon]^{-1}\Upsilon^{T}=(\psi)^{-1}\Upsilon^{T} \tag{4.1}$$

The ψ tends to become singular whose non-singularity can be ensured by decreasing the number of hidden layer neurons or increasing number of training data N in the initialization phase.

$$\hat{\beta}=[\Upsilon^{T}\Upsilon]^{-1}\Upsilon^{T}\Gamma \tag{4.2}$$

Let us suppose that we have a set of initial training set $(x_i,t_i), 1\leq i\leq N_0$. The ELM learning for this data presents the problem of minimizing error $||\Upsilon_0\beta-\Gamma_0||$ for

$N_0 \geq L$ where

$$\Upsilon_0 = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_L, b_L, x_1) \\ \vdots & \cdots & \vdots \\ G(a_1, b_1, x_{N_0}) & \cdots & G(a_L, b_L, x_{N_0}) \end{bmatrix}_{N_0 \times L} \quad \text{and} \quad \Gamma_0 = \begin{bmatrix} t_1^T \\ \vdots \\ t_{N_0}^T \end{bmatrix}_{N_0 \times m}$$

A solution to minimizing $||\Upsilon_0 \beta - \Gamma_0||$ is equal to $\hat{\beta}^{(0)} = (\psi_0)^{-1} \Upsilon_0^T \Gamma$. Suppose that we obtain another set of training data that contains N_1 number of observations

$$\Upsilon_1 = \begin{bmatrix} G(a_1, b_1, x_{N_0+1}) & \cdots & G(a_L, b_L, x_{N_0+1}) \\ \vdots & \cdots & \vdots \\ G(a_1, b_1, x_{N_0+N_1}) & \cdots & G(a_L, b_L, x_{N_0+N_1}) \end{bmatrix}_{N_1 \times L} \quad \text{and} \quad \Gamma_1 = \begin{bmatrix} t_{N_0+1}^T \\ \vdots \\ t_{N_0+N_1}^T \end{bmatrix}_{N_1 \times m}$$

Now the error minimization problem is transformed into two sets of data

$$\epsilon = \left|\left| \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix} \beta - \begin{bmatrix} \Gamma_0 \\ \Gamma_1 \end{bmatrix} \right|\right| \tag{4.3}$$

Gathering information from both sets of data, the output weight matrix $\beta^{(1)}$ becomes

$$\beta^{(1)} = (\psi_1)^{-1} \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix}^T \begin{bmatrix} \Gamma_0 \\ \Gamma_1 \end{bmatrix} \tag{4.4}$$

$$\psi_1 = \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix}^T \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix} = \begin{bmatrix} \Upsilon_0^T & \Upsilon_1^T \end{bmatrix} \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix} = \psi_0 + \Upsilon_1^T \Upsilon_1 \tag{4.5}$$

For sequential learning, we are required to represent $\beta^{(1)}$ as a function of $\beta^{(0)}$,ψ_1, Υ_1 and Γ_1

$$\begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix}^T \begin{bmatrix} \Gamma_0 \\ \Gamma_1 \end{bmatrix} = \Upsilon_0^T \Gamma_0 + \Upsilon_1^T \Gamma_1 = \psi_0 \psi_0^{-1} \Upsilon_0^T \Gamma_0 + \Upsilon_1^T \Gamma_1$$

$$= \psi_0 \beta^{(0)} + \Upsilon_1^T \Gamma_1 = \left(\psi_1 - \Upsilon_1^T \Upsilon_1\right) \beta^{(0)} + \Upsilon_1^T \Gamma_1 = \psi_1 \beta^{(0)} - \Upsilon_1^T \Upsilon_1 \beta^{(0)} + \Upsilon_1^T \Gamma_1 \tag{4.6}$$

From Equations 4.4 and 4.6

$$\beta^{(1)} = (\psi_1)^{-1} \begin{bmatrix} \Upsilon_0 \\ \Upsilon_1 \end{bmatrix}^T \begin{bmatrix} \Gamma_0 \\ \Gamma_1 \end{bmatrix} = (\psi_1)^{-1} \left[\psi_1 \beta^{(0)} - \Upsilon_1^T \Upsilon_1 \beta^{(0)} + \Upsilon_1^T \Gamma_1\right]$$

$$= \beta^{(0)} + (\psi_1)^{-1}\Upsilon_1^T\left[\Gamma_1 - \Upsilon_1\beta^{(0)}\right] \tag{4.7}$$

For the $k+1$ set of training data, where $k \geq 0$ and N_{k+1} corresponds to the number of observations

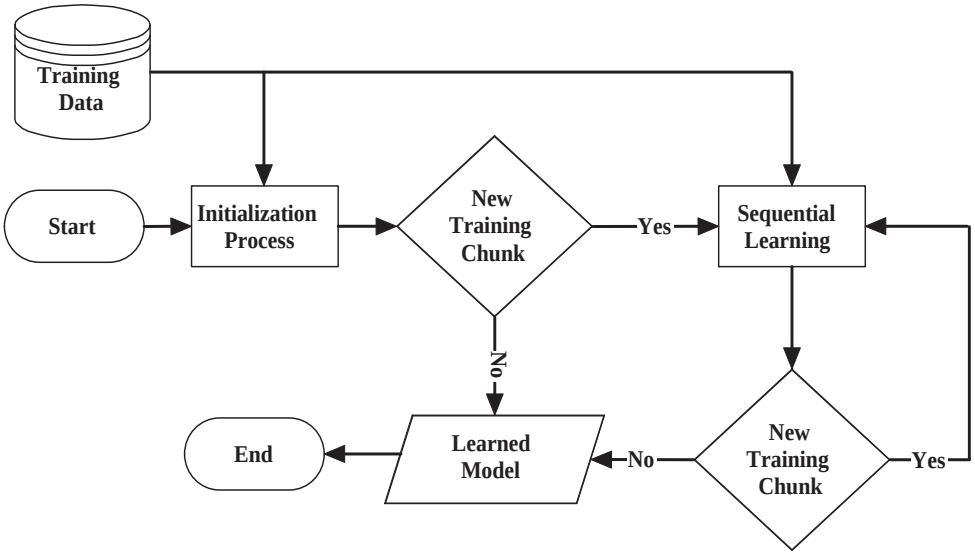


Figure 4.7: Data flow diagrams for recursive learning strategy.

$$\Upsilon_{k+1} = \begin{bmatrix} G(a_1, b_1, x_{(\sum_{j=0}^{k} N_j)+1}) & \cdots & G(a_L, b_L, x_{(\sum_{j=0}^{k} N_j)+1}) \\ \vdots & \cdots & \vdots \\ G(a_1, b_1, x_{(\sum_{j=0}^{k+1} N_j)}) & \cdots & G(a_L, b_L, x_{(\sum_{j=0}^{k+1} N_j)}) \end{bmatrix}_{N_{k+1}\times L},$$

$$\Gamma_{k+1} = \begin{bmatrix} t^T_{\sum_{j=0}^{k}(N_j)+1} \\ \vdots \\ t^T_{\sum_{j=0}^{k+1}(N_j)} \end{bmatrix}_{N_{k+1}\times m}.$$

By generalizing the arguments from Equations 4.5 and 4.7, a recursive approach for updating the least-square solution is

$$\psi_{k+1} = \psi_k + \Upsilon_{k+1}^T\Upsilon_{k+1} \tag{4.8}$$

$$\beta^{k+1} = \beta^k + \psi_{k+1}^{-1}\Upsilon_{k+1}^T(\Gamma_{k+1} - \Upsilon_{k+1}\beta^k) \tag{4.9}$$

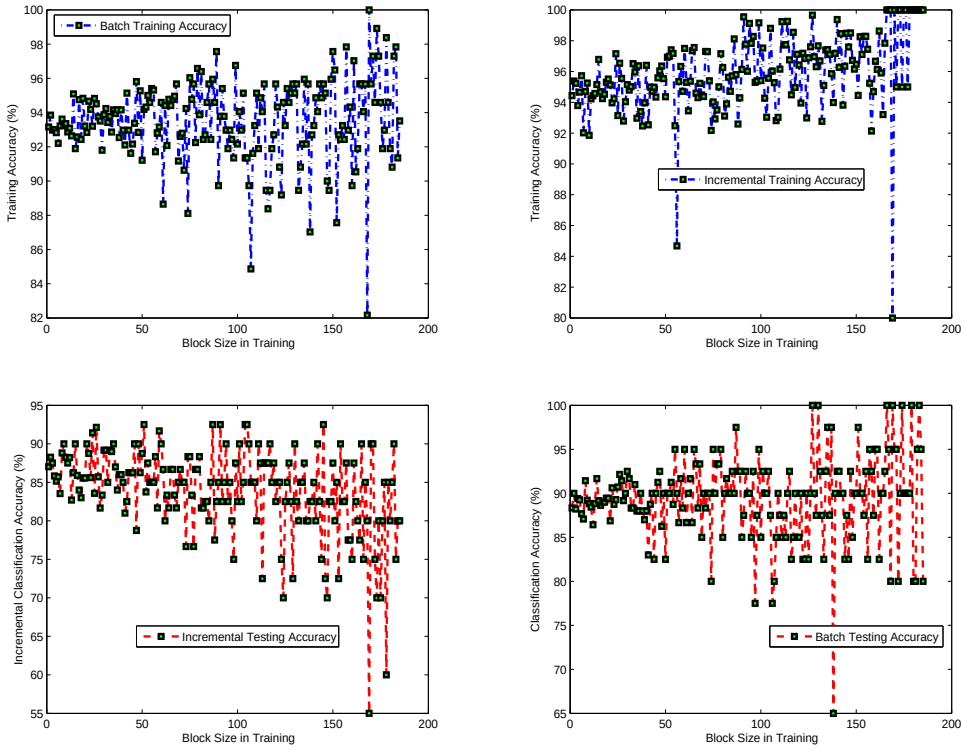


Figure 4.8: Performance comparison of batch mode learning vs. recursive scheme for ELM using concatenated PHOG features of blocks for randomly selected *actions* from Weizmman dataset. Training (top row) and testing accuracy (bottom row) are shown.

$$\psi_{k+1}^{-1} = \left[\psi_k + \Upsilon_{k+1}^T \Upsilon_{k+1}\right]^{-1} = \psi_k^{-1} - \psi_k^{-1}\Upsilon_{k+1}^T \left[I + \Upsilon_{k+1}\psi_k^{-1}\Upsilon_{k+1}^T\right]^{-1} \tag{4.10}$$

The new derivation achieves a similar learning performance as we attain using a traditional ELM, subject to the condition $rank(\Upsilon_0) = L$. It should be noted that the sizes of incoming training sets need not be equal. The recursive learning approach presented in Equations 4.8-4.10 consists of two stages - *1)* initialization and *2)* sequential learning phases. During the former stage, Υ_0 is prepared with the condition that the number of distinct training samples should be equal to or greater than the number of hidden neurons. Later, the sequential learning phase updates the model by receiving new training data in a one-by-one or group-by-group fashion. Figure 4.7

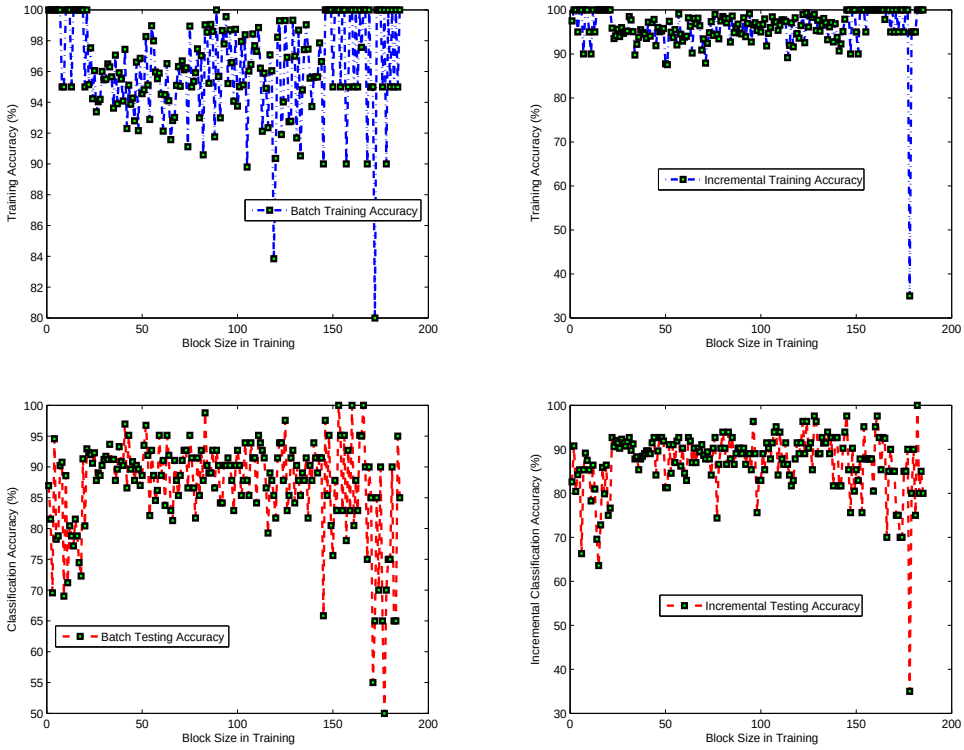


Figure 4.9: Performance comparison of batch mode learning vs. recursive scheme for ELM using PHOG features of track window for randomly selected *actions* from Weizmman dataset . Training (top row) and testing accuracy (bottom row) are shown.

represents a flow diagram of the recursive strategy of ELM. It is worth mentioning that the recursive performance (presented above) becomes equal to traditional ELM learning if all the training samples are provided in initialization phase, i.e., $N_0 = N$. The only control parameter to be selected for modified ELM is the size of network i.e. L. Clearly, training a classifier requires large number of training samples to ensure better performance during the test phase. The recursive edition of ELM (Equations 4.8-4.10) allows us to sequentially update our model for the new set of training data. Apparently, the presented recursive learning scheme offers a simple analytic solution to its batch mode counterpart. However, it is important to analyze the performance

degradation because of accumulated errors and higher order of computations during model update. We perform a number of rigorous tests using Weizmman dataset to judge the performance of online training. The results presented are compared against traditional ELM working in batch learning mode. Figures 4.8-4.9 show the accuracy acquired both in the training and testing phase using traditional ELM (left columns) and recursive ELM (right column).

The performance comparison of concatenated PHOG features obtained from blocks is shown in Figure 4.8 whereas a similar experiment is repeated using PHOG features of a track window in Figure 4.9 while using 90% of data for learning a model. For incremental learning the data is provided in the form of sets each consisting of ten training samples while the number of hidden neurons is also equal to the size of samples in individual training sets. The majority of available data is used in learning as an attempt to investigate the performance variations based on training module which is evidently a fresh component in an old framework. We achieve a difference in classification using both kinds of ELMs within a range of less than 5% thus confirming the claim (Equations 4.8-4.10) that the performance achieved by using sequentially learnable scheme is comparable to the batch mode learning. Furthermore, another set of experiments employing same actions and types of features is performed where the amount of training data is gradually changed from 10% to 90% to explore the behavior of sequential learning when the amount of data varies from too small to very large (Figure 4.10). It is to be noted that the accuracy of both schemes using varying percentage of training features is very close. However, the serial learning performs relatively better because of smaller approximation errors accumulated during training update for each incoming set of training data. The size of training sets does not influence recognition whereas an overall decrease in training time is observed on increasing the size of training sets which is a plausible phenomena since larger groups of data lead to a decrease in number of times the model requires to be updated. It is worth noting that the training and testing accuracy of recursive scheme is, respec-

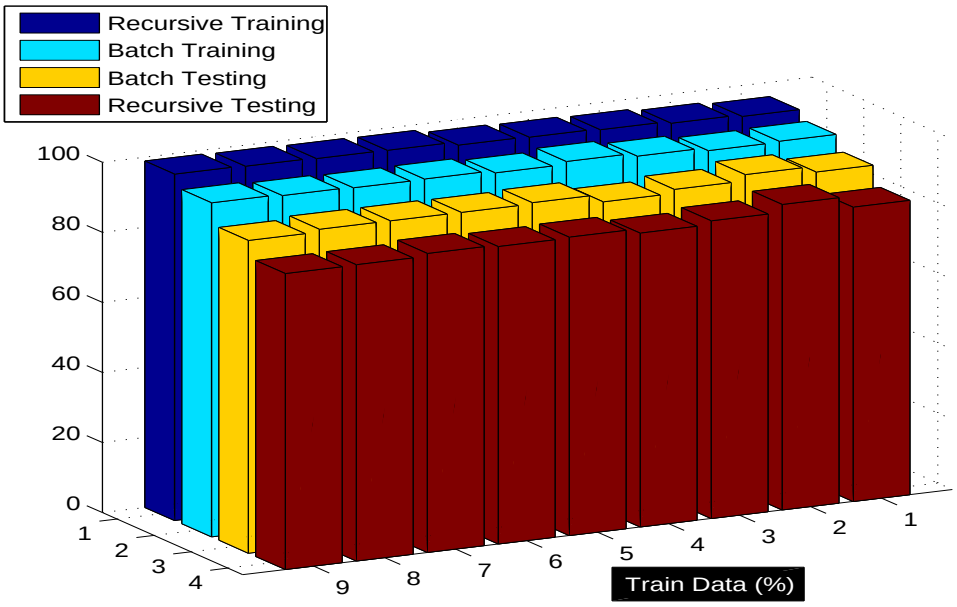


Figure 4.10: For concatenated PHOG features of blocks; performance comparison of batch mode learning vs. recursive scheme for ELM using varying percentage of training data for randomly selected *actions* from Weizmman dataset.

tively, higher and lower than batch mode learning which shows its tendency of slight over-fitting however, such small degradation is still acceptable owing to its practicality for real-life applications.

4.3 Results and Discussions

Before providing accuracy analysis of our proposed framework; a diminutive introduction to our experimental setup is presented including details of used datasets along with feature extraction and classifier architecture. In terms of required preprocessing, our method needs the least attention mechanism as compared to other schemes which solely rely on specialized steps such as video trimming, foreground masks and a fixed size bounding box centered at the person of interest. Our method belongs to a class which favors sparse sampled features, hence, in the end only four feature vectors with

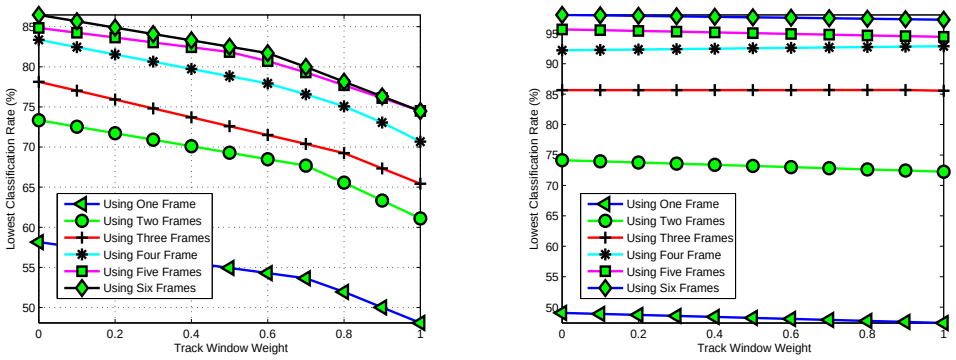


Figure 4.11: Lowest classification for changing number of frames from Weizmman dataset using (*Left*) separate PHOG features from blocks (*right*) concatenated PHOG features of blocks.

the maximal coverage of a moving human body are extracted. Regarding dataset, we use Weizmman set which has become a de-facto standard for human action recognition. There is no established testing protocol and various methods use different number of training samples, feature dimensions and learning parameters for their respective classifiers; comparisons presented in this dissertation are the best quoted results. However, it should not be counted as direct comparison due to significant dissimilarities among architectures of different frameworks. For simple classification, a bank of binary ELMs is applied to classify each individual class of the actions. An alternative option could be the use of multiclass ELM which requires normalized data, however, the associated computational overhead for such schemes out weighs the improvement. At the same time requirement of distinctive set of training samples critically influences the classification performance as well. The final classification in our proposed scheme is based on combining individual estimates of features from track window and blocks using weights w_i and $(1 - w_i)$, respectively. Keeping in view the overall details of our proposed framework, the first and foremost concern to address is the effect of concatenation of PHOG feature vectors extracted from blocks ς_i. Another important aspect is the effect of utilizing features from multiple

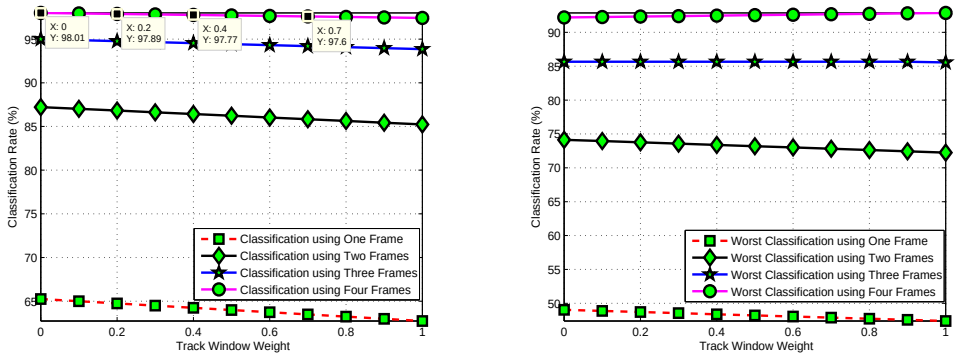


Figure 4.12: Using changing number of frames from Weizmman dataset *Left*: Accuracy analysis and, *right*: Worst classification analysis.

frames to labels an action. Intuitively, concatenation of features produces a vector of larger length which, perceptibly, offers distinctive information because of its relation to a higher dimensional feature space. Please refer to Figure 4.11 which represents the lower classification achieved using Leave-One-Out cross validation for Weizmman set. It is noticeable that the lower end of achieved accuracy approximately varies between $5 - 10\%$ for PHOG feature vectors of ς_i on using varying number of frames without (*left* column) and with concatenation (*right* column). The graph lines for multiple frames are obtained using video frames from current and past time instances only and as the number of frames increases the recognition accuracy also improves whereas a one-to-one correspondence emerges among features vectors of blocks and respective video frames because of concatenation. As observed in Figure 4.6, the feature vectors for track windows of different actions have higher proximity thus their less discriminative information promotes the idea to assign them lower weights in final classification. It is evident from results in Figures 4.11 for un-concatenated features that assigning higher weight to track window degrades the performance of our final classification. The classification renders stable path on using an increasing number of frames and concatenated feature vectors of blocks (Figure 4.12) where we achieve 98.01% accuracy by using only four frames compared with the use of entire video for

a similar level of recognition [28]. It is noteworthy contribution that the recognitions presented in Figures 4.11-4.14 are for classification of *snippets* not the whole videos. Instead of completely eliminating track window information, we use them as harmonizing feature vectors for reliable recognition since occasionally they outperform features emanating from rectangular blocks ς_i. Based on these results, our method

Table 4.1: Classification comparison against different approaches at *snippet* level.

	Our Method			**[46]**	**[18]**	**[16]**	**[17]**		
Frames Used	1/1	3/3	6/6	10/10	1/12	1/9	1/1	7/7	10/10
Accuracy (%)	65.2%	95.0%	99.63%	99.6%	55.0%	93.8%	93.5%	96.6%	99.6%

closely falls into the problem domain of Schindler's method [17]; Table 4.1 presents the accuracy analysis of various methods operating on a collection of frames, with or without look-ahead assumption of a video sequence. It is clear that for *snippet* of length 1, the performance of [17] is the best but with the rising length of *snippet* our method, utilizing lowest number of frames, outperforms all other frameworks. The

Table 4.2: Classification comparison of different methods at sequence level.

Our Method	**[46]**	**[18]**	**[16]**	**[17]**	**[28]**	**[1]**	**[55]**	**[54]**
100.0%	100.0%	72.8%	98.8%	100.0%	97.8%	92.6%	99.44%	95.04%

effect of changing number of frames and assigning weights to track window is better realized in Figures 4.13-4.14; the accuracy becomes more stable and achieves close to perfect recognition using *snippet* of only 6 frames as compared to 10 frames required by [46] and [17] for the similar accuracy (see Table 4.1).

Our method is not designed for a video level classification, however, for better comparison we present the recognition analysis using majority voting of *snippets* of length 1 of proposed method against other schemes (Table 4.2). This comparison validates the claim that short collection of video frames after resourceful processing is

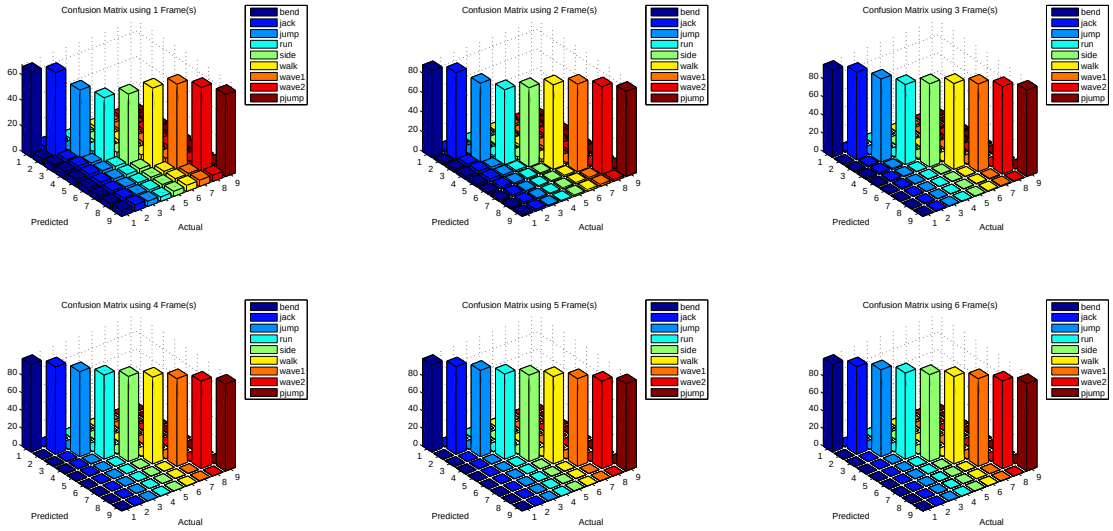


Figure 4.13: Confusion matrices for varying number of frames 1-6 (*left-right* and *top-bottom*)of videos taken from Weizmman dataset (this figure is best viewed in colors).

almost as informative as entire video to label an action. Our method achieves perfect classification to produces comparable results against [46],[17].

4.4 Summary

A method for action recognition is presented which uses adaptively extracted PHOG features from full body and its annotated parts based on tracking strategy that utilizes both shape and appearance information. The fundamental motivation behind this work is to recognize an action using a small collection of video frames which can serve as building blocks of an action. The detailed experimental evaluation confirms the enhanced performance of adaptively extracting PHOG features whereas the idea can easily be extended to actions detection of multiple humans in a video frame. Furthermore, it has been shown that the method performs well against state-of-the-art methods using as low as six frames and without impractical assumptions such as

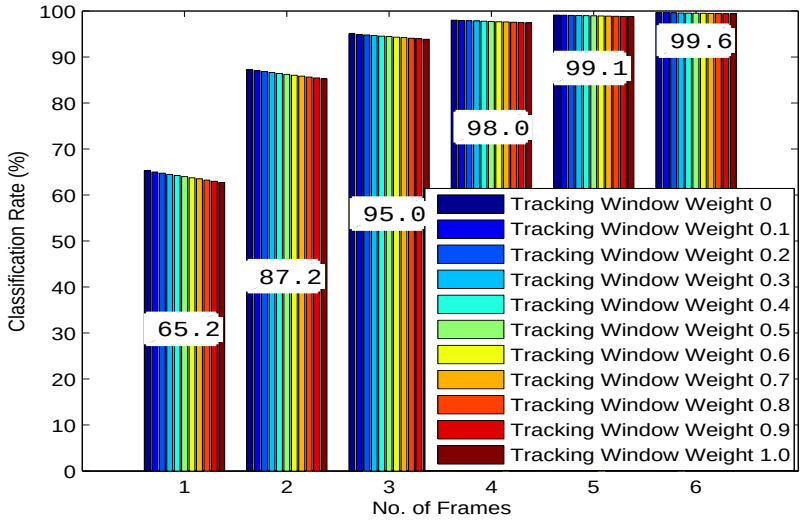


Figure 4.14: Recognition analysis of proposed method for varying number of frames and weight of track window (this figure is best viewed in colors).

look-ahead, video stabilization and available foreground masks. The use of *snippets* also conforms to reliable recognition for videos containing numerous actions which are not identifiable by the sequence level schemes.

Chapter 5

Conclusion and Future Work

5.1 Concluding Remarks

This dissertation investigates the learning of multiple types of features for visual recognition. For still images, global analysis is utilized in an attempt to preserve the correlation amongst neighboring pixels which provides compact and reliable content representation. The global analysis of videos requires volumetric processing to decrease computational load and extract features which are repeatedly identifiable in the temporal domain. The spatio-temporal feature sets are efficiently computed because of shift-invariance and motion selectivity properties of the 3D DT-CWT. These features support reduced artifacts, better localization and resourceful processing of a video. The computed spatio-temporal features have shown an enhanced ability to capture the dynamics of various actions performed by similar or different objects. To reduce the dimensionality of features obtained through global analysis, bidirectional 2D-PCA is proposed which greatly minimizes not only the number of coefficients but also conserves the correlation in orthogonal directions.

To add complementary information, sets of raw local features are extracted from input data and further quantized into representative sets utilizing unique pruning strategies. To obtain flexible representations of objects instances visual vocabularies

are employed since a target undergoes large shape variations while performing an action. For available feature vectors, learning a model is the trickiest and the most expensive phase because of the requirement of careful selection of various parameters. To simplify and minimize computational load of training, ELM is applied which can analytically perform the learning operation with improved generalization and minimized error.

Despite the successfulness of visual vocabularies which are highly sensitive to the size of a cluster, they lack spatial information within their representations. Spatial and geometric information plays an important role in a recognition framework. In addition, batch mode learning also restricts the application of traditional frameworks to various fields where new categories are frequently introduced into a system. To overcome these shortcomings, the framework proposed in earlier part of this dissertation is refined based on tracking using shape and appearance information to extract PHOG features. Conventionally, PHOG features in recognition are computed for a specified region of interest (ROI). Such ROIs in this scheme are identified by adaptive representation of a moving human body using an effective tracking scheme. Finally, a recursive least-square solution to determine an output matrix of the hidden layer of neurons is applied. The recursive least-square solution supports online learning which eliminates the need to relearn whole data if a new chunk of training data is received.

5.2 Future Work

The frameworks proposed in this dissertation can be extended in a variety of ways such as overcoming the limitations or extending its applications to new domains such as crowd analysis, surveillance and automotive industry.

5.2.1 Overcoming Limitations

The limitations of current framework is the unavailability of mechanism to deal with view point changes. Also, it is still an open problem to declare which *snippets* are

the most discriminative and contain key information to recognize an action. Such information can help to decide length and combining spaced out frames in *snippets* for recognizing a particular action. Practically there are two uninvestigated but fundamental questions in compute vision research; what are the building block units of an action and how complicated can actions become? Further investigation is needed to improve the tracking part by embedding prediction-correction mechanism which can greatly reduce computational cost while trying to locate the most probable location of the track window.

5.2.2 Onset Prediction of Critical Events

The prediction of the beginning of an abnormal behavior in a crowd, a collection of dynamic objects, is a relatively new direction of research in computer vision. The proposed recognition framework can be extended towards onset prediction of critical events or abnormal behavior. Potential applications of this research include industry, military, weather forecasting, traffic monitoring, surveillance and sports events.

5.2.3 Nonlinear Theory of Behavior Analysis in Recognition

In the field of automated recognition and classification there is a growing need to apply psychological methods to analyze the behavior of dynamic objects. Studies show that the thinking abilities and the response to a similar situation varies significantly for different species. Such changes in behavior are based on physical maturity, strength, and whether a particular object is alone or in a group. Initially, one can apply the holistic approach for prediction at the start of an abnormal crowd behavior because the existing correlation in a crowd is not taken into consideration. It is believed that the nonlinear dynamics of interactions amongst group of humans are based on different parameters which can be modeled using the nonlinear theory of behavioral analysis. For behavior classification, the idea of ELM can be extended to the temporal

domain since classification of an observed behavior without considering correlation among adjacent frames may deteriorate recognition.

References

[1] S. Ali, A. Basharat, M. Shah, "Chaotic Invariants for Human Action Recognition", Proc. of the IEEE Conf. on CV, pp. 1–8, 2007.

[2] M.J. Black, "Explaining Optical Flow Events with Parameterized Spatio-temporal Models", Proc. of the IEEE Conf. on CVPR, pp. 1326–1332, 1999.

[3] M. Brand, N. Oliver, A. Pentland, "Coupled HMM for Complex Action Recognition", Proc. of the IEEE Conf. on CVPR, pp. 994–999, 1997.

[4] T.J. Burns, "A Non-homogeneous Wavelet Multiresolution Analysis and Its Application to the Analysis of Motion", PhD thesis, Air Force Institute of Tech., 1993.

[5] A.A. Efros, A.C. Berg, G. Mori, J. Malik, "Recognizing Actions at Distance", Proc. of the IEEE Conf. on CV, 2003.

[6] L. Gorelick, M. Blank, E. Shechtman, M. Irani, Basri R. "Actions as Space Time Shapes", IEEE Trans on PAMI, pp. 2247–2253, 2007.

[7] G.B. Huang, Q.Y. Zhu, C.K. Siew, "Extreme Learning Machine: Theory and Applications", Neurocomputing, pp. 489–501, 2005.

[8] H. Jiang, M.S. Drew, Z.N. Li, "Successive Convex Matching for Action Detection", Proc. of the IEEE Conf. on CVPR, pp. 1646–1653, 2006.

[9] N.G. Kingsbury, "Complex Wavelets for Shift Invariant Analysis and Filtering of Signals", Journal of Applied and Computational Harmonic Analysis, vol. 10, no. 3, pp. 234–253, 2001.

[10] D.G. Lowe, "Distinctive Image Features from Scale-invariant Key Points", International Journal of Computer Vision, 60(2), pp. 91-110, 2004.

[11] I. Laptev, "On Space-time Interest Points", International Journal of Computer Vision, 64(2/3), pp. 107-123, 2005.

[12] J. Liu, S. Ali, M. Shah, "Recognizing Human Actions Using Multiple Features", Proc. of the IEEE Conf. on CVPR, pp. 1–8, 2008.

[13] J. Liu, J. Luo, M. Shah, "Recognizing Realistic Actions from Video "in the Wild"", Proc. of the IEEE Conf. on CVPR, 2009.

[14] J.M. Morel, G.Yu, "ASIFT: A New Framework for Fully Affine Invariant Image Comparison", SIAM Journal on Imaging Sciences, vol. 2, issue 2, 2009.

[15] G. Mori, X. Ren, A.A. Efros, J. Malik, "Recovering Human Body Configurations: Combining Segmentation and Recognition", Proc. of the IEEE Conf. on CVPR, pp. 326–333, 2004.

[16] H. Jhuang, T. Serre, L. Wolf, T. Poggio, "A Biologically Inspired System for Action Recognition", Proc. of the IEEE Conf. on CV, 2007.

[17] K. Schindler, L.v. Gool, "Action Snippets: How Many Frames Does Human Action Recognition Require?", Proc. of the IEEE Conf. on CVPR, 2008.

[18] J. Niebels, F.F. Li, "A Hierarchical Model of Shape and Appearance for Human Action Classification", Proc. of the IEEE Conf. on CVPR, pp. 1–8, 2007.

[19] I.W. Selesnick, K.Y. Li, "Video Denoising Using 2D and 3D Dual-tree Complex Wavelet Transforms", Wavelet Applications in Signal and Image Proc., SPIE 5207, San Diego, 2003.

[20] I.W. Selesnick, F. Shi, "Video Denoising Using Oriented Complex Wavelet Transforms", Proc. of the IEEE Int. Conf. on Acoust., Speech, and Signal Proc., (ICASSP), vol. 2, pp. 949–952, 2004.

[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, vol. 6, pp. 123–151, 2005 (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 Trans. on PAMI, vol., no. 1, pp. 131–137, 2004.

[24] Y. Pang, D. Tao, Y. Yuan, X. Li, "Binary Two-Dimensional PCA", IEEE Trans. on Systems Man and Cybernatics - Part B, vol. 38, no. 4, pp. 1176–1180, 2008.

[25] A. Yilmaz, M. Shah, "Actions Sketch: A Novel Action Representation", Proc. of the IEEE Conf. on CVPR, pp. 984–989, 2005.

[26] Y. Wang, G. Mori, "Max-Margin Hidden Conditional Random Fields for Human Action Recognition", Proc. of the IEEE Conf. on CVPR, 2009.

[27] J. Yuan, Z. Liu, Y. Wu, "Discriminative Subvolume Search for Efficient Action Detection", Proc. of the IEEE Conf on CVPR, 2009.

[28] L. Wang, D. Sutter, "Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model", Proc. of Int. Conf. on CVPR, 2007.

[29] J. Niebles, H. Wang, F.F. Li, "Unsupervised Learning of Human Action Categories Using Spatio-temporal Words", Proc. of BMVC, 2006.

[30] P. Dollár, V. Rabaud, G. Cotteral, S. Belongie, "Behavior Recognition via Sparse Spatio-temporal Features", In Proc. 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, pp. 2450–2454, 2009.

[32] C. Schüldt, I. Lepatev, B. Caputo, "Recognizing Human Actions: A Local SVM Approach", Prof. of Int. Conf. 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", Proc. of the IEEE Conf on CVPR, 2008.

[35] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning Realistic Human Actions from Movies", Proc. of the IEEE Conf on CVPR, 2008.

[36] D. Xu, S. Yan, D. Tao, L. Zhang, X. Li, H-J. Zhang, "Human Gait Recognition with Matrix Representation", IEEE Trans. on Circuits and Systems for Video Technology, vol. 16, no. 7, 896–903, 2006.

[37] R. Yu, A. Baradarani, "Sampled-data Design of FIR dual Filter Banks for Dual-tree Complex Wavelet Transforms", IEEE Trans. on Signal Proc., vol. 56, no. 7, pp. 3369–3375, 2008.

[38] U.v. Luxburg, "A Tutorial on Spectral Clustering", Statistics and Computing, 17(4), 2007.

[39] W-Y. Chen, Y. Song, H, Bai, C-J. Lin, E.Y. Chang, "Parallel Spectral Clustering in Distributed Systems", anonymous

[40] M. Leordeanu, M. Hebert, "A Spectral Clustering for Corresponding Problem Using Pairwise Constraint", Proc. of ICCV, vol. 2, pp. 1482–1489, 2005.

[41] E. Shechtman, M. Irani, "Space-time Behavior Based Correlation", In Proc. of the IEEE Conf. on CVPR, 2005.

[42] G. Cheung, S. Baker, T. Kanade, "Shape-from-silhouette of Articulated Objects and Its Use for Human Body Kinematics Estimation and Motion Capture", Proc. of the IEEE Conf. on CVPR, 2003.

[43] A. Bobick, J. Davis, "Recognition of Human Movement Using Temporal Templates", IEEE Trans. on PAMI, 23(3), pp. 257–267, 2001.

[44] D. Weinland, R. Ronfard, E. Boyer, "Free Viewpoint Action Recognition Using Motion History Volumes", Computer Vision and Image Understanding, 104(2-3), pp. 249–257, 2006.

[45] S. Wong, T. Kim, R. Cipolla, "Learning Motion Categories Using Both Semantics and Structural Information", Proc. of the IEEE Conf. on CVPR, 2007.

[46] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, "Actions as Space-Time Shapes", Proc. of the IEEE Conf. on CV, 2005.

[47] J. Liu, "Learning Semantic Features for Visual Recognition", PhD Thesis, University of Central Florida, pp.1-175, 2009.

[48] A. Baradarani, Q.M. Jonathan Wu, "Moving Object Segmentation Using the 9/7-10/8 Dual-tree Complex Filter Bank, Proc. of the 19th IEEE ICPR, 2008.

[49] M. Varma, A. Zisserman, “A Statistical Approach to Texture Classification from Single Images”, International Journal of Computer Vision, vol.62(1–2), pp. 61–81, 2005.

[50] B.A. Olshausen, D.J. Field, “Sparse Coding With an Over-complete Basis Set: A Strategy Employed by v1?”, Vision Research, vol. 37, pp. 3311-3325, 2007.

[51] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis”, Machine learning, vol. 42, pp. 177–196, 2001.

[52] R. Minhas, M. Abdul Adeel, Q.M. Jonathan Wu, “A Fast Recognition Framework Based on Extreme Learning Machine Using Hybrid Object Information”, pp. 1831–1839, Neurocomputing (73), 2010.

[53] Y. Boykov, M.-P. Jolly, “Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Object in n-d Images”, Proc. of ICCV, pp. 105–112, 2001

[54] R. Minhas, A. Baradarani, S. Sepideh, Q.M. Jonathan Wu, “Human Action Recognition Using Non-separable Oriented 3D Dual-tree Complex Wavelet Transform”, Proc. of Asian Conference on Computer Vision, 2009.

[55] R. Minhas, A. Baradarani, S. Sepideh, Q.M. Jonathan Wu, “Human Action Recognition Using Extreme Learning Machine Based on Visual Vocabularies”, pp. 1906–1917, Neurocomputing (73), 2010.

[56] J. Shi, C. Tomasi, “Good Features to Track”, Proc. of the IEEE Conf. on CVPR, 1994.

[57] A. Bosch, A. Zisserman, X. Munoz, ”Representing Shape with a Spatial Pyramid Kernel”, Proc. of Int. Conf. on Image and Video Retrieval, 2007.

[58] J. Ponce, J. Brady, "Towards a Surface Primal Sketch", Three Dimensional Machine Vision, pp. 195-240, 1987.

[59] R.C. Bolles, R. Horaud, "3DPO: A Three Dimensional Part Orientation System", Kluwer Academic, 1987.

[60] D.G. Lowe, "The Viewpoint Consistency Constraints", International Journal of Computer Vision, vol. 1(1), pp. 57–72, 1987.

[61] J. Huang, S.R. Kumar, M. Mitra, W. Zhu, R. Zabih, "Image Indexing Using Color Correlograms", Proc. of the IEEE Conf. on CVPR, pp. 762-768, 1997.

[62] W.T. Freeman, E.H. Adelson, "The Design and Use of Steerable Filters", IEEE Trans. on PAMI, vol. 13, pp. 891–906, 1991.

[63] M. Urban,J. Matas, O. Chum, T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremum Regions", In Proc. of BMVC, 2002.

[64] K. Mikolajczyk, C. Schmid, "Scale and Affine Invariant Interest Point Detectors", International Journal of Computer Vision, pp. 63–86, 2004.

[65] K. Grauman, T. Darrell, "Pyramid Match Kernels: Discriminative Classification with Sets of Image Features", In Proc. of ICCV, 2005.

[66] A. Oliva, A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope", International Journal of Computer Vision, pp. 145–175, 2001.

[67] F. Lv, R. Nevatia, "Single View Human Action Recognition Using Key Pose Matching and Viterbi Path Searching", Proc. of the IEEE Conf. on CVPR, 2007.

[68] A. Bissacco, M.H. Yang, S. Soatto, "Detecting Humans via Their Pose", Proc. of the NIPS Conf., 2007.

[69] A. Johnson, M. Hebert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes, IEEE Trans. On PAMI, Vol. 21, No.5, 1999.

[70] Y. LeCun, L. Bottou, G.B. orr, K.-R. M̈uller, "Efficient Backprop", Lecture Notes in Computer Science, vol. 1524, pp. 9–50, 1998.

[71] N.-Y. Liang, G.-B. Huang, P. Saratchandran, N. Sundrarajan, "A Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks", IEEE Trans. on Neural Networks, vol. 17, pp. 1411–1423, 2006.

[72] S.M.S. Nejhum, J. Ho, M.-H. Yang, "Visual Tracking with Histogram and Articulated Blocks", Proc. of the IEEE Conf. on CVPR, 2008.

[73] Li Fei-Fei, R. Fergus, A. Torralba, "Recognizing and Learning Object Categories", Short Course at ICCV, 2009.

[74] R. Fergus, P. Perona, A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning", Proc. of the IEEE Conf. on CVPR, pp. 264–272, 2003.

[75] M. Weber, M. Welling, P. Perona, "Unsupervised Learning of Models for Recognition", Proc. of 6th ECCV, pp. 18–32, 2000.

[76] P. Felzenszwalb, D. Huttenlocher, "Pictorial Structures for Object Recognition", International Journal of Computer Vision, 61(1), pp. 55–79, 2004.

[77] P. Viola, M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features", Proc. the IEEE Conf. on CVPR, pp. 511–518, 2001.

[78] S. Agarwal, D. Roth, "Learning Sparse Representation for Object Detection", Proc. of ECCV, pp. 113–130, 2002.

[79] B. Leibe, A. Leonardis, B. Schiele, "Combined Object Categorization and Segmentation with an Implicit Shape Model", Proc. ECCV Workshop SL in C. Vision, pp. 17–32, 2004.

[80] P. Viola, M. Jones, D. Snow, "Detecting Pedestrians Using Patterns and Motion Appearance", International Journal of Computer Vision, pp. 734–741, 2003.

[81] G. Y. Dorko, C. Schmid, "Selection of Scale-Invariant Parts for Object Class Recognition", Proc. of ICCV, pp. 634–640, 2003.

[82] H. Schneiderman, T. Kanade, "Object Detection Using the Statistics of Parts", International Journal of Computer Vision, 56(3), pp. 151–177, 2004.

[83] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, 60(2), pp. 91–110, 2004.

[84] K. Mikolajczyk, C. Shmid, "An Affine Invariant Interest Point Detector", Proc. of ECCV, pp. 128–142, 2002.

[85] M. Ozuysal, V. Lepetit, P. Fua, "Fast Key Point Recognition Using Random Ferns", IEEE Trans. on PAMI, (2009) To appear.

[86] A. Opelt, M. Fussenegger, A. Pinz, P. Auer, "Weak Hypothesis and Boosting for Generic Object Detection and Recognition", Proc. of ECCV, pp. 71–84, 2004.

[87] S. Ali, M. Shah, "A Supervised Learning Framework for Generic Object Detection in Images", Proc. of ICCV, pp. 1347–1354, 2005.

[88] G-B. Huang, Q-Y. Zhu, C-K. Siew, "Extreme Learning Machine: Theory and Applications", Neurocomputing, pp. 489–501, 2005.

[89] R. Zhang, G-B. Huang, N. Sundarajan and P. Saratchandran, "Multicategory Classification Using an ELM for Gene Expression for Cancer Diagnosis", Computational Biology, pp. 485–495, 2007.

[90] Q-J. Benedict, S. Emmanuel, "ELM for Classification of Music Genres", Proc. ICARCV, pp. 1–6, 2006.

[91] G-B. Huang, H.A. Babri, "Upper bound on number of hidden neurons in F.N. with arbitrary bounded nonlinear activation functions, Neural Networks", pp. 224–229, 1998.

[92] J. Yang, David Zhang, F. Frangi, J-Y. Yang, "Two-Dimensional PCA: A New Approach to Appearance Based Face Representation and Recognition", IEEE Trans on PAMI, 26(1), pp. 131–137, 2004.

[93] P. Sanguansat, W. Asdornwised, S. Marukatat, S. Jitapunkul, "Two-Dimensional Random Subspace Analysis for Face Recognition", Proc. of ISCIT, pp. 628–631, 2007.

[94] D. Zhang, Z-H. Zhou, $(2D)^2$PCA for Efficient Face Representation and Recognition", anonymous.

[95] R. Minhas, A.A. Mohammed, Q.M. Jonthan Wu, "A Generic Moments Invariant Based Supervised Learning Framework for Classification Using Partial Object Information", Proc. of Conf. on CRV, Canada, pp. 45–52, 2009.

[96] A. Opelt, A. Pinz, A. Zisserman, "Incremental Learning of Object Detectors Using a Visual Shape Alphabet", Proc. of the IEEE Conf. on CVPR, pp. 3–10, 2006.

[97] J. Shotton, A. Blake, R. Cippola, "Contour-Based Learning for Object Detection", Proc. ICCV, pp. 503–510, 2005.

[98] A. Torralba, K.P. Murphy, W.T. Freeman, "Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection", Proc. of the IEEE Conf. on CVPR, pp. 762–769, 2004.

[99] D. Tao, X. Li, X. Wu, S.J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gair Recognition", IEEE Trans on PAMI, 29(10), pp. 1700–1715, 2007.

[100] J. Sun, D. Tao, S. Papadimitriou, P.S. Yu, C. Faloutsos, "Incremental Tensor Analysis: Theory and Applications", ACM Trans on Knowledge Discovery from Data, 2(3), pp. 11:1–11:37, 2008.

[101] D. Tao, X. Li, X. wu, W. Hu, S.J. Maybank, "Supervised Tensor Learning, Knowledge and Information Systems", pp. 13:1–42, 2007.

[102] K. Mikolajczyk, C. Schmid, "A Performance Evaluation of Local Descriptors", IEEE Trans on PAMI, 27(10), pp. 1615–1630, 2005.

[103] A. Bosch, A. Zisserman, X. Munoz, "Representing shape with a spatial pyramid kernel", Int. Conf. on Image and Video Retrieval, 2007.

[104] A. Opelt, A. Pinz, A.Zisserman, "A Boundary Fragment Model for Object Detection", In Proc. of ECCV, pp. 575-588, 2006.

[105] R. Minhas, A. Baradarani, S. Sepideh, Q.M. Jonathan Wu, "Human Action Recognition Using Extreme Learning Machine via Multiple Types of Features", Symposium on SPIE Defense, Security, and Sensing , 2010 (Accepted).

[106] M.A. Fischler, R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Comm. Assoc. Comp. Mach., pp. 381–395, 1981.

[107] W.E.L. Grimson, T. Lozano-Perez, "Localizing Overlapping Parts by Searching the Interpretation Tree", IEEE Trans. on PAMI, pp. 469–482, 1987.

[108] C. Schmid, R. Mohr, "Local Gray Value Invariants for Image Retrieval", IEEE Trans. on PAMI, 19(5), pp. :530-536, 1997.

[109] N. Dalal, B. Triggs, "Histograms of Oriented Gradients for Human Detection", Proc. of the IEEE Conf. on CVPR, 2005.

[110] J. Winn, A. Criminisi, T. Minka, "Object Categorization by Learned Universal Visual Dictionary", Proc. of ICCV, 2005.

[111] H. Moravec, "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover", In Tech Report CMU-RI-TR-3 Carnegie-Mellon University, Robotics Institute, 1980.

[112] S.M. Smith, J.M. Brady, "Susan - A New Approach to Low Level Image Processing", International Journal of Computer Vision, 23(1), pp. :45–78, 1997.

[113] K. Mikolajczyk, C. Schmid, "Scale and Affine Invariant Interest Point Detectors", International Journal of Computer Vision, 60(1), pp. 63–86, 2004.

[114] D. Ziou, S. Tabbone, "Edge Detection Techniques an Overview", International Journal of Pattern Recognition and Image Analysis, 8(4), pp. 537–559, 1998.

[115] T. Lindeberg, "Edge Detection and Ridge Detection with Automatic Scale Selection", International Journal of Computer Vision, 30(2), pp. 117–154, 1998.

[116] C. Harris, M. Stephens, "A Combined Corner and Edge Detector", In Proc. of the Aley Vision Conference, pp. 147–151, 1988.

[117] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, "Groups of Adjacent Contour Segments for Object Detection", IEEE Trans. on PAMI, pp. 36–51, 2008.

[118] K. Yan, R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors", In Proc. of the IEEE Conf. on CVPR, pp. 506–513, 2004.

[119] H. Ning, W. Xu, Y. Gong, T.S. Huang, "Latent Pose Estimator for Continuous Action Recognition:, Proc. of the ECCV, 2008.

[120] H. Ning, Y. Hu, T.S. Huang "Discriminative Learning of Visual Words for 3D Human Pose Estimation", Proc. of the IEEE Conf. on CVPR, 2008.

[121] C. Fanti, L. Zelnik-Manor, P. Perona, "Hybrid Models for Human Recognition", Proc. of ICCV, 2005.

[122] F. Porikli, "Integral Histogram: A Fast Way to Extract Histogram in Cartesian Spaces", In Proc. of the IEEE Conf. on CVPR, pp. 829–836, 2005.

[123] D. Freedman, T. Zhang, "Interactive Graph Cuts Based Segmentation with Shape Priors", In Proc. of the IEEE Conf. on CVPR, pp. 755–762, 2005.

[124] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories", Proc. of the IEEE Conf. on CVPR, 2006.

Vita Auctoris

Rashid Minhas was born in a small district of Pakistan called Layyah. He obtained his Bachelor of Science (with computer science/mathematics) degree from Bahauddin Zakariya University Multan, Pakistan in 1997. From 1997 to 2004 he worked with various IT departments in capacity as software developer and network administrator. Rashid moved over to Gwangju Institute of Science and Technology (GIST) for Master of Science degree in Mechatronics Engineering to avail a prestigious IITA scholarship offered by the Ministry of Information and Communication, Republic of Korea where he completed his studies with the noteworthy first position in Departmental Annual Mechatronics Project Competition for the year 2004.

In 2006, he entered PhD program at the department of ECE of the University of Windsor, Canada. Rashid won doctoral tuition scholarship, Fredrick Atkins graduate award, and finalist for the best paper award at ICAL'08 during his graduate studies at UWindsor. In recognition of excellent performance in course work and research, after a university wide competition he has been selected for the post-doctoral fellowship of Ministry of Research and Innovation, ON, Canada.