Silhouette based Human Action Recognition

To be submitted as Thesis in partial fulfilment of the requirement for the degree of

Master of technology

In

Microwave & Optical Communication

Submitted By

Awantika Dwivedi

(2k15/MOC/06)

Delhi Technological University, New Delhi, India

Under The Supervision of

Dr. Rajeev Kapoor

Department of Electronic and Communication

Delhi Technological University

(Formerly Delhi College of Engineering)

Shabad Daulatpur, Main Bhawana Road

New delhi-110042, India

DECLARATION

I Awantika Dwivedi, hereby declare that the work entitled **"Silhouette based Human Action Recognition"** *has been carried out by me under the guidance of Dr. Rajiv Kapoor, in Delhi Technological University, New Delhi.*

This project is a part of the degree of M.Tech in Microwave & Optical Communication. This is an original work and all sources; reference and literature used and excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Awantika Dwivedi
2K15/MOC/06

CERTIFICATE

This is to certify that the dissertation entitled: **Silhouette based Human Action Recognition** *in the partial fulfilment of the requirements for the reward of the degree of Masters of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by her under my guidance. The information and data enclosed in this project is original.*

Dr. Rajiv Kapoor
(Professor)
Department of Electronic and Communication
Delhi Technological University
(Formerly Delhi College of Engineering)

(Signature & Seal of Head of Department)

ACKNOWLEDGEMENT

I would like to express my gratitude to **Delhi Technological University, New Delhi** *for giving me opportunity to be a part of it. I would like to thank* **Dr. S. Indu** *Head of Department for her kind leadership. Most of all I am very highly obliged to my project advisor* **Dr. Rajiv Kapoor, Professor, Department of Electronic and Communication**, *Delhi Technological University, New Delhi for his exemplary guidance, monitoring and constant encouragement. I would also like to thank him for sparing the efforts in compiling the work presented here. I would also like to thank Mr. Om Mishra, my parents and my friends for giving me guidance and support.*

Awantika Dwivedi

2K15/MOC/06

ABSTRACT

Human motion analysis is currently receiving increasing attention from computer vision researchers. This interest is motivated by applications over a wide spectrum of topics. For example, segmenting the parts of the human body in an image, tracking the movement of joints over an image sequence, and recovering the underlying 3D body structure are particularly useful for analysis of athletic performance, as well as medical diagnostics. The capability to automatically monitor human activities using computers in security-sensitive areas such as airports, border crossings, and building lobbies is of great interest to the police and military. With the development of digital libraries, the ability to automatically interpret video sequences will save tremendous human effort in sorting and retrieving images or video sequences using content-based queries. Other applications include building man-machine user interfaces and video conferencing.

The research trend in the field of action recognition has recently led to more robust techniques, which to some extent are applicable for action recognition in complex scenes. Action recognition in complex scenes is an extremely difficult task due to challenges such as background clutter, camera motion, occlusions and illumination variations. To address these challenges, several methods, like tree-based template matching, tensor canonical correlation, prototype based action matching, incremental discriminant analysis of canonical correlation, latent pose estimation and a generalised Hough transform were proposed. Most of these methods are very complex and require pre-processing, like segmentation, tree data structure building, target tracking, background subtraction or the fitting of a human body model. On the other hand, recently, spatio-temporal features have gained popularity because of their state-of-the-art performance with reduced or even no pre-processing. These methods apply interest point detectors and local descriptors to characterize and encode the video data, and thereby perform action classification.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Overview

The human visual framework has a mind blowing ability in quick and precise examination and translation of visual information. We are capable not exclusively to distinguish and perceive static items, be that as it may, likewise to recognize examples of movement and fleeting occasions in complex situations. Indeed, even a youngster can essentially indicate whether a man is strolling, running or lighting. Over the most recent couple of decades, the strengthening of machines with the capacity to decipher visual information has seen huge improvement, yet the best in class is no place close what the human visual framework can accomplish. The corpus of video information is developing quickly. All of the regularly utilized customer equipment, for example, portable PCs, cell phones or computerized cameras can record recordings. Because of video sharing and long range interpersonal communication sites a tremendous measure of video information is being transferred and shared ordinary. This tremendous measure of information requires brilliant devices for efficient seeking and route. Right now the seeking and recovery depend on literary labels. However issues like disgraceful labeling or different dialects more often than not prompt recovery of disconnected material. Additionally, a few information are indescribable by words. All these propose that we utilize the more normal inquiry utilizing content which requires instruments for comprehension video information and deciphering it. With the expansion in reconnaissance frameworks, the requirement for programmed checking will develop. For example, a walk 2011 insights uncover that there is a sum of 1.85 million CCTV *cameras in UK. The internet preparing of this tremendous sum of the information stream is inconceivable for human work. In this manner programmed translation of video is an option.*

The essential concentration in this postulation is the programmed acknowledgment of human activity classes in groupings of the video. Specifically, given recordings containing a human performing an activity, we are keen on arranging the activity into one of the known classes or classes (e.g. running, applauding or boxing) which have been educated. Perceiving the conduct of people in arrangements of video can help in a wide assortment of utilizations including content-base video investigation, reconnaissance, human-PC connection, stride acknowledgment, what's more, virtual activity union. It is trying because of the expansive changes that exist inside one activity classification. For example, the activities should be possible by different people with differing appearances playing out the activity with different styles. Likewise, activities can occur under different conditions, for example, different enlightenments, foundations, perspectives or rates.

The terms activity and movement have been utilized conversely in the writing. To clear up their implications, we allude that denes activity as basic examples of movement performed by a solitary individual which normally goes on for a brief term. Cases of activities are bowing, running, waving and tossing. The movement then again is alluded to a grouping of activities performed by a few people in which the subjects may communicate with each other. Exercises are regularly done in any longer terms analyzed to activities. Cases are two people moving tango, a b-ball group scoring an objective or a bank assault. In our work, we focus on activities as opposed to exercises. All things considered, activities can be considered as the building pieces of exercises.

1.2 Motivations

Aside from scientific interest, our inspiration originates from the wide assortment of uses which could benefit from this PC vision framework. A portion of essential ones are [1,2]:

- *Content-based Video Retrieval and Summarization - Video sharing sites like YouTube, Google Video, Vimeo, and Daily motion require more ancient calculations for Presentation looking and recovery of the video content. Albeit looking in light of content might be quicker in an accumulation of a huge number of recordings, they may not generally recover recordings that are identified with the questioned video. A characteristic option is to dissect the sight and sound substance of video and to recover in light of the hidden semantics. Utilizing procedures for human activity acknowledgment can be beneficial towards translating the substance and explanation of recordings. So also rundown is another imperative application particularly for a substance like game occasions or news.*

- *Behavioural Intelligence - The lion's share some portion of the security and reconnaissance frameworks working in airplane terminals, banks, shopping centres or healing facilities depend on camcorders, which generally need human administrators to recognize any anomalous exercises. By expanding utilization of these frameworks, PC based answers for observing to supplant or help human administrators are required.*

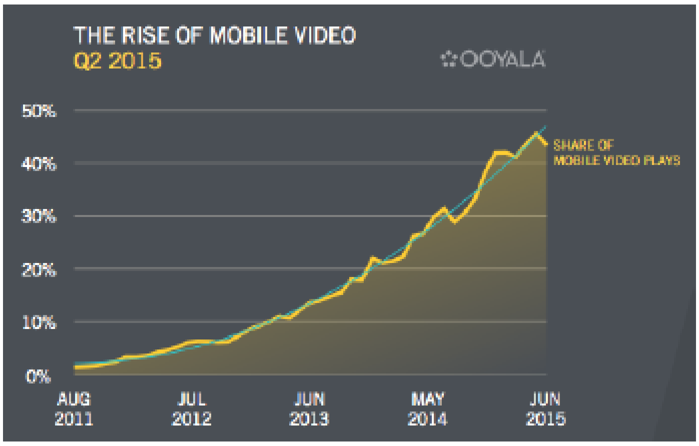


Fig 1.1. Rise in access of videos through years. Credits- OOYALA

- *Human-PC Interaction - Most machines get inputs by means of interfaces such as console, mouse and touch screen. In spite of the fact that these gadgets can cooperate with high accuracy and lower cost, they are less characteristic and expressive contrasted with utilizing signals. Utilizing a magnet [3] or infrared LED [4] may not be constantly helpful and lovely furthermore, may compel development. Consequently, the more regular and engaging route is to depend on total visual data and attempt to perceive hand and body motions caught by the camera. So, human activity acknowledgment can assume a significant part towards more characteristic human-machine communications.*
- *Step Recognition - Biometrics manages the calculations for acknowledgment of human subjects from their physical or behavioural prompts. The previous incorporates fingerprints, confront, iris and so on. Perceiving subjects in light of physical biometrics require the collaboration of subjects also, is impossible now and again. So, behavioural biometrics have increased much consideration as of late which concentrates the acknowledgment of people from their conduct or style of doing their exercises. The benefit of these strategies is that subject's*

participation is most certainly not essential and it can continue without intrusion or meddle with the subject's movement.

- *Virtual Action Synthesis - Combining reasonable people or human movements in a virtual condition has wide applications in diversion industry and additionally the new liveliness. So any calculation in displaying the human movement can propel the business in these fields.*

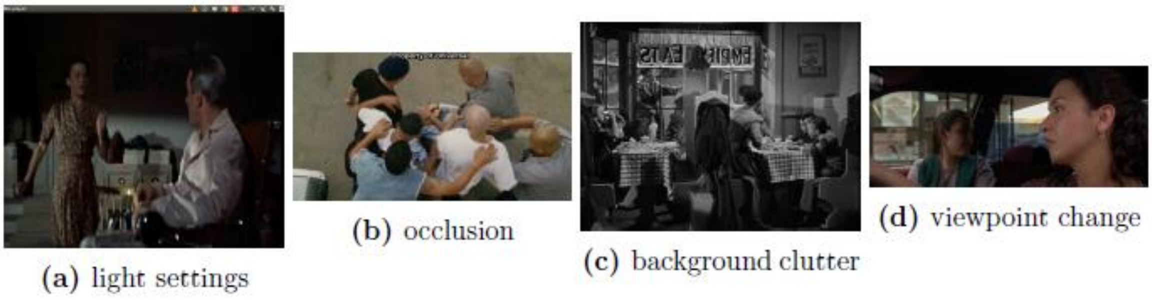
1.3 Challenges

In spite of extraordinary endeavours in the PC, vision looks into, unconstrained acknowledgment of human activities in true conditions stays unsolved as it were. This issue is trying since a PC based acknowledgment framework, learned on a restricted preparing set, needs to perceive an activity performed by an obscure individual in a jumbled foundation with ominous recording setting. Here we talk about the fundamental difficulties we confront in acknowledgment of human activities:

- *Intra-class Variance and Inter-class Similarity - Improvement of a programmed activity acknowledgment framework is an imposing test, mostly because of the conceivable varieties of an activity having a place with a similar class. Different subjects can play out the same activity. Their different appearance or garments change the elements removed from the activities, prompting blunders. Also, the style with which a man executes an activity may change, for example, there are different examples and walk lengths for strolling. A perfect activity acknowledgment framework ought, to sum up over fluctuations of a class keeping in mind the end goal to perceive a formerly inconspicuous case. Also, there might be significant closeness between examples of two different classes of* *activities. For instance strolling, running and running all have practically similar examples. Additionally, check watch and cross arms have similar developments somewhat. A decent activity acknowledgment framework ought to be ciently discriminant crosswise over different classes.*

❖ *Complex Background - The jumbled foundation may include commotion what's more, irrelevant points of interest to the framework which can prompt blunders in the acknowledgment. Point directions are typically difficult to track in jumbled foundations. Additionally, in the cases where outline elements are required, complex foundations make it difficult to extricate solid forefronts.*

❖ *Brightening Change - The enlightenment conditions in the video can change due to different variables including different edges of produced light, numerous wellsprings of lighting, re section from articles and camera immersion. The adjustment in the lighting conditions can change the separated components drastically, along these lines makes it difficult to perceive precisely.*



(a) light settings

(b) occlusion

(c) background clutter

(d) viewpoint change

Fig 1.2 Challenges in Computer Vision

- *Viewpoint Change - The adjustment in the camera perspective will change the picture perception because of camera viewpoint acts. This will bring about significant changes in movement and structure highlights. A decent activity acknowledgment framework ought to be vigorous to perspective change to some degree.*

- *Occlusion – Occlusion can be brought about by the subject itself (self-occlusion), by other objects or due to being outside the field of view. They prompt debasement of the visual appearance of parts of the subject. These parts might be basic or discriminative in perceiving an activity.*

- *Execution Rate Variance -There can be significant variety in the rate in which the activity is being executed. This can affect the activity acknowledgment particularly when the utilized components depend on movement or worldly windows. A powerful activity acknowledgment framework ought to be invariant to rates of execution of an activity. The accessibility of standard databases for activity acknowledgment has empowered a target*

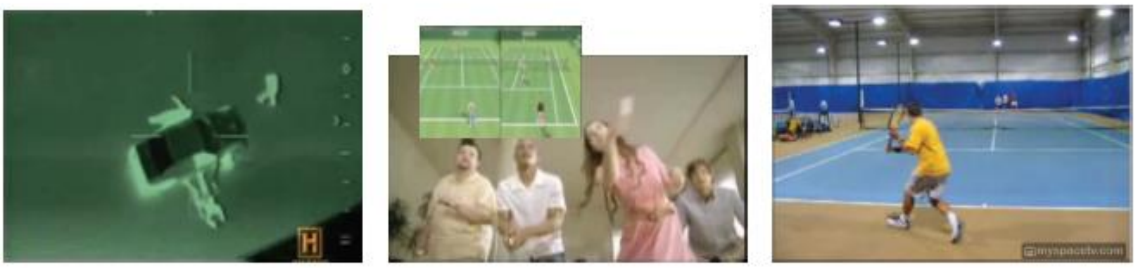
1.4 Contribution

Human activity acknowledgment in PC vision has pulled in solid research enthusiasm for late years as a result of its promising applications. Computerized discovery of a man's nuclear developments is called human activity acknowledgment. Assorted *techniques have been proposed for recognizing the human activity because of the elements in the basic portrayal of human body. This paper gives a look over the noteworthy investigates led around there and a thorough study of the distinctive scientific classifications recommended by various specialists. A framework is proposed for effective activity acknowledgment in view of the idea ‗bag of related postures' by assessing the relationship between's consecutive stances in an activity on both straight and non direct element vectors.*

1.5 Application Areas

Human movement comprehension can serve numerous application zones, running from visual observation to human PC interaction(HCI) frameworks. Especially, the application spaces are restricted to those that include camera setups. The following is a case rundown of such frameworks.

- *Visual Surveillance: As the video innovation turn out to be more typical, visual observation frameworks attempted a quick advancement prepare, and have more or, on the other hand less turn into a piece of our every day lives. Figure demonstrates a case observation infra-red (IR) video yield. Human activity comprehension can discover fraudful occasions –such as robberies, fightings, and so forth –, to identify pedestrains from moving vehicles, and can serve to track patients who require exceptional consideration (like distinguishing a falling individual [5]).*



(a) surveillance (b) home entertainment (c) sport annotation

Fig 1.3 Various applications of computer vision

- *Human-Computer Interaction: Ubiquitous registering has expanded the nearness of HCI frameworks all over the place. An as of late developing string is in the territory of electronic diversions and home entertainment(see Figure 2.2(b)). These frameworks are as of now in view of extremely innocent video and flag handling. In any case, as the innovation advance, the pattern will move towards more astute and complex HCI frameworks which include action and conduct understanding.*

- *Sign Language Recognition: Gesture acknowledgment, which is a subdomain of activity acknowledgment that works over the abdominal area parts, serves a considerable measure for programmed comprehension of gesture based communication [6,7,8].*

- *News, Movie and Personal Video Archives: By the lessening in the cost of video catching gadgets and by the advancement of sharing sites, recordings progress* *toward becoming to be a significant piece of the today's close to home visual chronicles. Programmed comments of those chronicles, together with film and news documents will offer assistance data recovery. Moreover, programmed explanation of news and game video archives(see Figure 2.2(c) for an illustration edge) is an essential string for getting to the important data in a brisk and simple way. Individuals might be intrigued in finding certain occasions, describable just by the exercises included and movement acknowledgment can help significantly for this situation.*

- *Social Evaluation of Movements: The perception of behavioral examples of people is very vital for the exploration of human science, engineering, and the sky is the limit from there. Machine view of exercises and examples can manage many examines in this zone. For instance, Yan et al. tries to discover assessments of where individuals invest energy by looking at head directions [9]. Strangely, look into like this one will help in urban arranging.*

1.6 Thesis Outline

The rest of the thesis is organised as mentioned: In chapter 2, we provide a brief overview of the primary and elementary terms that form the basis to understand the action recognition.

In chapter 3, we discuss the evolution of the method deployed here in relation with the other papers and methodologies i.e. related work or literature survey. In chapter 4, we describe our method by explaining the features like STIP, then, bag of correlated poses and dimensionality reduction technique like PCA, ICA and ISOMAP. In chapter 4, we *show the experimental results through various statistical data, graphs and confusion matrix. In chapter 5, we conclude and display the future work.*

Chapter 2

RELATED WORK

Visual analysis of human conduct has pulled in an awesome arrangement of consideration for the PC vision in view of the broad assortment of potential applications. The conduct of human being could be sectioned in atomic actions, every one shows a solo elementary movement. Perceiving action of human [10,11] is considered to be a segment in numerous applications, HCI, mainly in video surveillance, and proactive figuring. To lessen human intrusion in the examination of human conduct, unsupervised learning might be more reasonable than supervised learning. Be that as it may, the multifarious temperament of a person's behaviour investigation makes unsupervised taking in a tricky undertaking.

The activity portrayal and acknowledgment is moderately older idea yet juvenile. Despite the fact that the analysts are contributing tremendous endeavors to propose their methodologies for activity or movement acknowledgment—actually this field is still not material in numerous imperative territories. Thusly, this extend focus on the real groupings on appearance based activity portrayal. Clearly without having hearty and more intelligent portrayals. So it will be hard to perceive activities in a sensible way. The most widely recognized strategies connected for portraying picture signals are silhouette and contours, motion optical flows, colour and textures, depth maps. Silhoutte and in addition edges what's more, shapes, are utilized to fit human body in *pictures in light of the fact that the greater part of the body posture data stays in its outline. Be that as it may, strategies utilizing edges depend on a background subtraction arrange as a result of the trouble of separating human outlines in complex situations.*

As of late, profundity signs have been incorporated into a few human posture acknowledgment frameworks in view of the profundity maps given by the multi-sensor KinectTM. The novel profundity portrayal offers almost three dimensional data from a modest sensor synchronized with RGB information. Dominant part of the human activity acknowledgment works take after two ways. The principal procedures include at first finding the question of intrigue e.g. a human, following it so that a depiction of how the question changes after some time can be framed, and afterward at last arranging the activity. A large number of the endeavors utilizing this approach are tried and assessed on unchallenging datasets, for example, KTH which highlights a solitary individual playing out a solitary activity in uncluttered setting. While following a protest separate one action from another is not useful in jumbled scene. Tracking through movement or, on the other hand frontal area division can be touchy and corrupt awkwardly when blunders happen in adjacent edges inside a video succession.

The other methodology evades the downsides of following strategies straightforwardly breaking down the movement designs inside the whole outline all through the video succession. Here strategies consider, for example, video groupings, processing fleeting formats constituting two segments; a motion energy image (MEI) and a motion history image (MHI). The MEI is a twofold portrayal of the movement happening for the arrangements of edges whereas the MHI has a dim scale force portrayal at places *where the changes made in the most recent time is made to be the lightest. This thought is considered as a superior thought for activity acknowledgment because it catches the route fit as a fiddle and movement develop over time.*

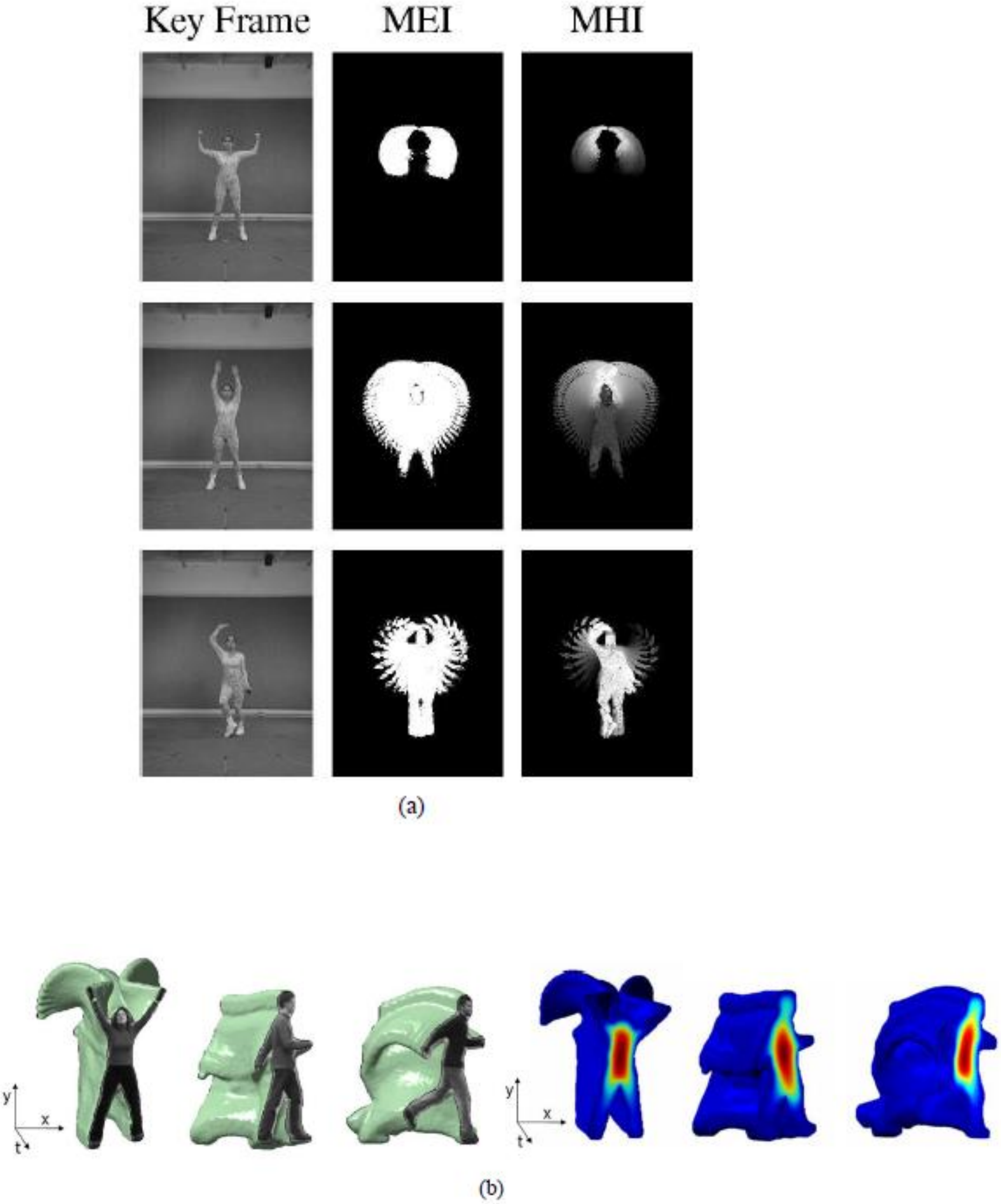


Fig 2.1(a) Bobick and Davis [Bobick 2001], MEI & MHI representation together with the key frames. (b) Blank et al.[Blank 2005] Spatio-temporal volume for a given sequence

In the second way single layered methodologies are very normal where a movement is thought to be a specific category of picture grouping and acknowledgment is performed over an obscure contribution by classifying it into its class. The analysis of human action sums up the techniques for single layered methodologies. It can be characterized into sequential approach and space time approach. Space time methodologies are those that perceive human exercises by breaking down the space time volumes of movement recordings. Also, the greater part of the movement acknowledgment strategies utilize spatialtemporal highlight descriptors. It also displays the activity of human as a specific 3-D volume in a space-time measurement or an arrangement of components are extricated from the volume. At that point volumes for video edges are developed by linking picture outlines along a period pivot.

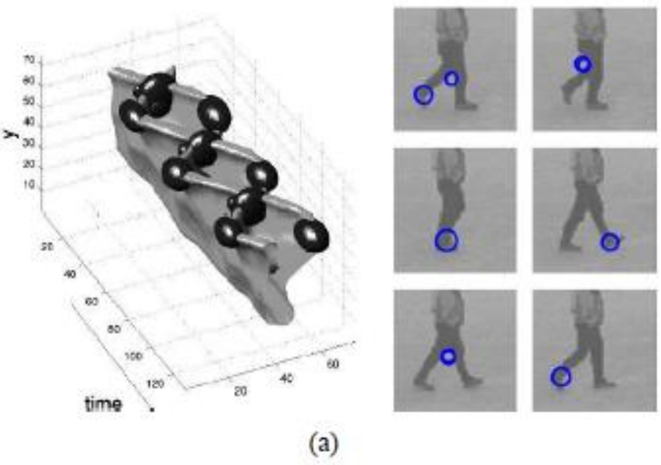


Fig 2.2 Laptev [Laptev 2005];Harris 3D interest point from the motion of leg of a walk

This is utilized for further similitude measures. Then again, sequential approaches regard human activities as a succession of specific perceptions. All the more particularly, they speak to a specific human activity as an arrangement of highlight vectors or descriptor separated from pictures and perceive exercises via looking for such a progression.

STIF based methodologies are accomplishing an expanding measure of consideration because of the dependability of the calculations under commotion and brightening changes. Be that as it may, these methodologies, in any case, can just create great outcomes under such circumstances where the datasets are generally less difficult in either constrained perspective changes or organized foundations. Notwithstanding that, the execution of highlight distance based classifier. Amongst the other strategies known Dimensionality reduction strategy is vital. Information starting from this present reality is regularly hard to comprehend on account of its elevated dimensionality.

The determination of the pictures might be 600x400 pixels, which implies that the info information has 240000 measurements. It stands to reason that a large portion of the order strategies endure and indeed, even bomb in their objectives when managing such sort of information because of their affectability to the dimensionality of the information . Clearly, that information winds up noticeably immovable from the computational perspective when long picture arrangements are utilized and a dimensionality diminishment system is required.

In order to attain a lower dimensional portrayal, a system of dimensionality diminishment is required after the human exercises are shown and recorded. Both, linear, (for example, PCA and ICA) and non-linear(for example, Isomap) dimensionality diminishment strategies can be connected to diminish the dimensionality of the information. The decision of either relies on upon the inherent way of the information set. Diminish the dimensionality of the information can be seen as a pre-preparing venture for arrangement purposes.

Different mechanized strategies are produced for recognizing action of humans. In the paper [10] the novel view dependent method to deal with the portrayal and acknowledgment of the movement by humans is introduced . These methodologies just pile the frontal area districts of a man (i.e., outlines) to monitor the change in shapes expressly. This strategy shows the portrayal of pictures utilizing temporal template – motion energy image (MEI) is the first esteem which demonstrates the existence of motion and motion history image (MHI) is the second esteem which is the corollary of sequence of motion. In the wake of figuring the different scaled MEIs and MHIs suggested strategy figure the Hu moments for every picture and then examine the Mahalanobis separation of the motion energy image framework for the familiar view sets. Any motion observed to be inside a limit separation is marked with action. This is thought to be a viable technique for speaking to and perceiving motion.

The paper [11] speak to activities as 3D shapes actuated by the outlines in the space-time volume. The strategy picks a Poisson-based descriptor since it reflects more global properties of the outline, and permits simple extraction of numerous helpful shape properties. It puts an effort to speak to a shape by depicting its outlines. The proposed technique has various focal points like: - it is potential to adapt to low quality video information, strategy does not require video arrangement, direct in the quantity of space-time focuses in the shape, and the general preparing time of technique takes under 30 seconds. This approach can likewise be connected with practically nothing change to general 3D shapes portrayal and coordinating. This technique is quick and vigorous, in light of the fact that it contains both the spatial data about the human posture, and the dynamic data, for example, global body motion and motion of the appendages in respect to the body. The local descriptors and holistic features are *encompassing elements accentuate distinctive parts of activities and are appropriate for the diverse sorts of activity databases.*

In this paper [12] a brought together activity acknowledgment structure melding holistic elements and local descriptors is introduced. Two dimensional and three dimensional SIFT which is a kind of local descriptor include descriptors in light of two dimensional SIFT focuses are separated what's more, all encompassing components separated with Zernike minutes. A combinational technique received here makes a practically identical outcome. The work done earlier, most nearby descriptors approaches utilize the spatiotemporal gradient data to concentrate intrigue interest point and, most holistic elements depend on the outlines or following. This paper acquaints outline differencing with concentrate over the holistic elements and local descriptors; after that utilization a bag of-words way to deal with feature vectors for the feature fusion.

In the paper [13] a new portrayal for human activities utilizing Correlogram of Body Poses (CBP) that forms favourable position of the temporal relationship of human stances and probabilistic appropriation. Silhouette is moderately simple to be gotten and it comprises of differentiated shape data, which is invariable with individual's gender orientation, body measure, illumination condition, apparel, and outlook. Normalized silhouette is utilized for contribution of framework because body postures have been determined by outlines, which is hearty to diverse attire, outlook, light variation and likewise it spares the calculation of description of features. The technique is hearty to lack of quality of the division, loud marking in preparing tests, and the pace of the activity.

An augmentation of Correlogram dependent posture identification is finished on [14]. Multiple sight activity acknowledgment utilizing local similarity random forest classifier *and multi-sensor combination is introduced in the thesis. The possibility of decision forest and Correlogram of body stance outlines since highlight descriptors is melded with various camera view on the IXMAS outline dataset. This type of combination technique would profit the acknowledgment exactness by depicting the activities with additional features. The issues like computational many-sided quality for clustering and dimensionality decrease can be illuminated with the randomized forest classifier. Crumbling because of the unequal execution of every camera concerning diverse activities is handled by another voting procedure. The irregular timberlands strategy has enhanced proficiency and viability over other learning calculations such as k-means particularly while managing vast scale information, and it will stay away from the over-fitting issue by fitting additional choice trees.*

The paper [15] has a shape-based element descriptor Pyramid Correlogram of Oriented Gradients (PCOG) which is computed from the MEI and MHI is utilized. After utilizing the nearby also, spatial design properties, the PCOG descriptor catches the fundamental data of human activities and gives great discriminative for order. Here another Human activity recognition system is proposed which focus on the human from the contribution by coordinating extricated HOG descriptors with the prototypical activity primitives. The gotten vectors are utilized for periodical activity apportioning. The yield just contains the district of intrigue (ROI). Once a finish practice cycle is extricated, two key edges and their comparing MHI and MEI are chosen to encode this development. At that point the activity class is anticipated by arranging the extricated highlight descriptors (PCOG) utilizing the prepared classifier utilizing the multi-class Support Vector.Machine (SVM) with the RBF bit. This methodology can recognize diverse activity classes from a video Sequence. An indoor condition, similar to a re *center is utilized for execution assessment. It is watched that meager portrayals in view of distinguished intrigue focuses, experience the ill effects of the loss of structure data .*

This paper [16] proposes a model which takes the movement and structure data at the same time and incorporates them in a bound together structure and hence gives an instructive what's more, minimal portrayal of human activities. By applying the movement layout to the volume with contrast of casing (DoF), the movement data is encoded into the movement highlight delineate (movement history picture), and structure include maps are acquired from the structure planes removed from the DoF volume. Two dimensional Gaussian pyramid and focus encompass operation are performed on each element outline, request to decay highlight maps into sub-band pictures restricted in numerous middle spatial frequencies. At that point naturally roused elements are separated utilizing a two-organize include extraction step ie; Gabor sifting and max pooling. The viability of this technique is assessed on the KTH, the multi-see IXMAS, and the UCF sports datasets.

Paper [17] proposes a programmed video comment calculation by incorporating semi-supervised learning and shared structure examination into a joint system for human activity acknowledgment. Another Semi-supervised Feature Correlation Mining (SFCM) strategy is presented which influences shared basic investigation for activity acknowledgment. Input training Action Videos may contain both named recordings and unlabeled recordings. Highlights extraction by Harris3D and Hoard/HoF BoW portrayal is performed for both preparing what's more, trying recordings to speak to them. As indicated by the conveyance of the visual components, a diagram model is developed in training. Expanding upon the chart, virtual marks of the unlabeled information can be created, amid which shared structural analysis is connected to reveal the element relationships to make the outcomes more dependable. Along these *lines, a classifier is prepared for activity acknowledgment. To assess exhibitions of proposed calculation a concise correlation is made with SVM with two kernel. Bayes Optimal Kernel Discriminant Analysis (BKDA), Taylor-Boost (T-Boost) and Semi-supervised Discriminant Analysis (SDA).*

Agarwal and Triggs portray [18] an approach for recuperating 3D human body posture from single pictures and monocular picture groupings. It identify human stance by direct nonlinear relapse against shape descriptor vectors extricated from outlines where an outline shape is encoded by histogram of-shape-settings descriptors. A sparse Bayesian approach is proposed in view of nonlinear regression algorithm called Relevance Vector Machine (RVM). The principle fascination of this technique is it doesn't require a 3D body model or marking of picture positions.Also, the strategy is effortlessly versatile to various individuals or appearances. This strategy indicate promising outcomes, being around three times more precise than closest neighbor strategies.

A similar approach has the distinctive execution on the distinctive database. This is because of the distinctive qualities of these databases. Some dataset has a bigger information scale, diverse situations, changing foundations because of the camera zoom, more people playing out a specific activity, and more intra-class uniqueness in the state of figures. Others may have a much lower information scale, just a single situation, static foundation, more activity classes and more between class comparability in the neighborhood movement. Taking after area gives brief prologue to the datasets utilized as a part of this for human activity acknowledgment which can be considered as benchmarks.

A. Weizmann Dataset

The Weizmann database contains grouping dataset and power dataset. The Weizmann order dataset is for activity preparing and grouping. The Weizmann power dataset is for trying the strength of a human acknowledgment technique and contains two information subsets. It contains 90 low-determination (180 × 144) video arrangements from nine individuals, each performing ten regular activities: run, walk, skip, hopping jack, twist, bounce set up on two legs, jogging sideways, wave one hand and wave two hands. Every one of the recordings are caught from a settled perspective.

B. KTH Dataset

This dataset was exhibited by _ and is an all the more troublesome dataset when diverged from the Weizmann dataset. The dataset involves six basic activities, specifically; Walking, Jogging, Running, Boxing, Hand-waving, and Hand-praising. Each action has 100 recordings for four remarkable circumstances in different light conditions, indoor and outdoors conditions. All these video groupings are recorded in a uniform establishment with a static camera of packaging rate 25fps and also down-inspected to the spatial assurance of 160x120 pixels. The recording conditions of the recordings in the KTH enlightening file are not stationary and there is a great deal of camera advancement and lighting impacts from time to time. The specimen edge of this dataset is appeared in fig 10. Along these lines, the framework extraction is not straight forward and clear establishment subtraction technique may not be sensible. From now on, for the outline extraction, a surface based division procedure is utilized.

C. IXMAS Dataset

Over two sets were recorded in controlled and rearranged settings. The primary practical activity dataset gathered from motion pictures and explained from motion *picture scripts is made in INRIA. INRIA Xmas Motion Acquisition Sequences (IXMAS). Multiview dataset utilized for view-invariant human activity acknowledgment. IXMAS expect to shape a dataset practically identical to the current best in class" in real life acknowledgment. It contains 11 activities, each performed 3 times by 10 on-screen characters of 5 guys or 5 females. The performing artists openly change their introduction for each securing with a specific end goal to show the view-invariance and no further signs on the most proficient method to play out the activities next to the marks.*

Chapter 3

METHODOLOGY

Figure 1 demonstrates essential strides and methodologies of recognition of human actions. A single action video is handled to create the grouping of the pictures. The Human Action Recognition procedure utilizes grouping of pictures. From grouping pictures silhoutte are created. From each casing of informational index, "Region of Interest" (ROI) comprising of the binary action shape is extricated.

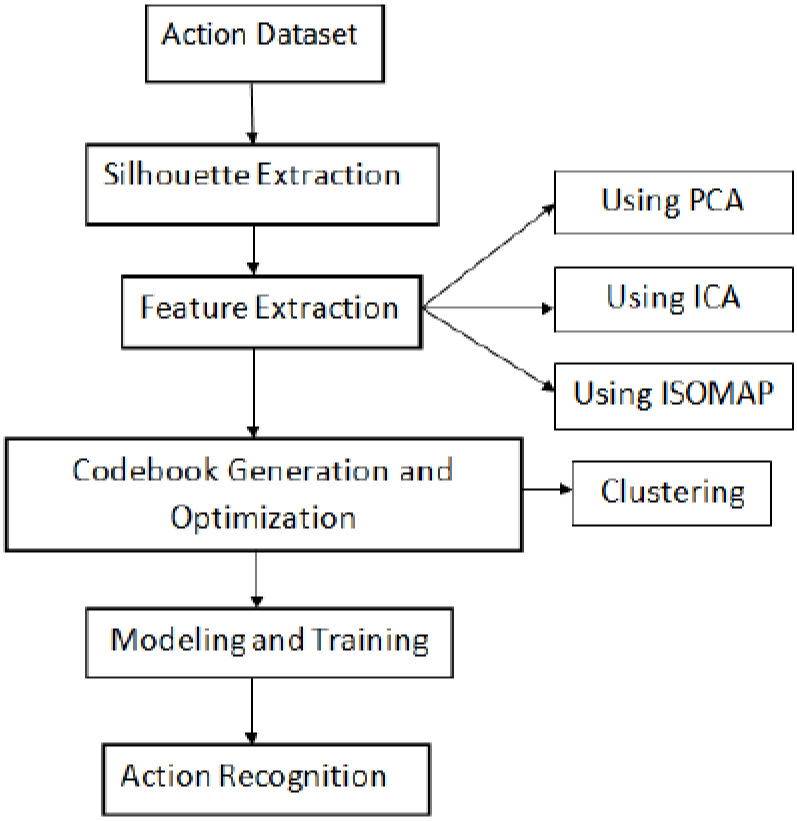


Fig 3.1 Framework of Human Action Recognition

Each video cut comprises of human involving only one action. Extraction of features is an exceptional type of dimensionality lessening. The input information is changed into the arrangement of fundamental feature vector is called feature extraction. From the separated features, human activities are recognized. Principal Component Analysis (PCA) is the prevalent technique for the feature extraction. Another technique is Independent Component Analysis (ICA). Linear Discriminate Analysis (LDA) is the arrangement device that can be connected on PC what's more, IC feature keeping in mind the end goal to show signs of improvement result. This work moreover incorporates an unsupervised non-linear strategy called ISOMAP. For the most part two vector quantization calculations are utilized: in particular customary K-mean clustering and Linde, Buzo, what's more, Gray (LBG)s clustering algorithm . At long last classification is finished utilizing the classifiers like SVM, multiSVM or HMM.

A. Bag-of-correlated poses

The pack of-correlated approach is an outstanding technique for activity acknowledgment. The pack of-components based methodologies can be connected in grouping by utilizing features as words. The Bag-of-Features portrayal is ordinarily a normalizes histogram, where each bin in the histogram is the number of features relegated to a specific code partitioned by the aggregate number of features in the video. Because of its prevalence, analysts are widely considering this system for their investigates. It has comparative forms as

- Bag-of-Feature
- Bag-of-Words
- Bag-of-Visual-Words
- Bag-of-Vocabularies
- Bag-of-Video-Words
- Bag-of-Points

The conventional bag-of-features portrayal dismisses structural data among the visual words. On the off chance that the codebook turns out to be extensive, it might deliver lower acknowledgment. To encode the structural data Correlogram of human stances in an activity succession is presented in [4]. Correlogram can also be taken as graphical *portrayal of autocorrelation. Bag of correlated stances is a generally new region of research, however a wide assortment of promising preferences are shown in this technique. Body postures encoded by silhouette are thought to be strong to various dress, appearance and light changes and it is the most ideal way to identify motion. The extricated normalized outlines are utilized as info elements for the Bag-of-Features (BoF) display.*

The proposed correlogram of body stances contains both temporal and statistical relationship data, which empowers the calculation to be fit for recognizing activities with comparable posture measurements however unique temporal ordering.

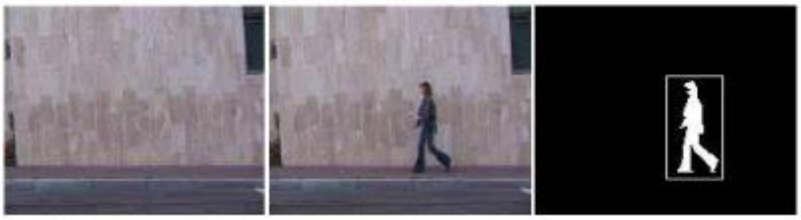


Fig 3.2 Detection of Region of Interest

In the interest point based activity acknowledgment technique as appeared in figure 3.2, each component vector is a 3-D descriptor figured around a distinguished intrigue point in an action sequence. In this strategy each feature vector is changed over from the 2-D outline mask to a 1-D vector by checking the mask from upper left to base right pixel by pixel.

Accordingly, every frame at the time t in an activity arrangement is spoken to as a vector of binary components, the length can be obtained as

M = row∗ column (1)

where ―row and ―column are measurements of the normalized stance outline. Assume the ith activity succession comprises of Si casings, then an activity succession can be spoken to as a network Bi with Pi rows and M segments.

Each column of the network remains for a solitary casing. Along these lines, for a preparation set with n activity successions, the entire preparing dataset can be spoken to as

$B = [B1;B2; . . . ;Bn]$ (2)

The aggregate number of rows, which is additionally the aggregate frame number in the preparation dataset, is

$P= P1 + P2 + . . . + Pn$ (3)

Since components are in high-dimensional space, we initially utilize PCA for dimensionality diminishment. Consequently, each edge Ft is anticipated into a lower measurement. At that point, visual vocabulary can be developed by grouping highlight vectors acquired from all the preparation tests utilizing the k-means. The measure of the visual vocabulary is the quantity of the cluster k. A shading correlogram is a 3-D framework where every component demonstrates the co-event of two hues those are at a certain separation from each other. In real life portrayal, every component in BoCP signifies the probabilistic co-occurrence of two body postures occurring at a specific time contrast from each other. Since the stances are separated into k groups, the dimensionality of the correlogram grid at a settled time balance Δt is k ∗ k, where k speaks to the codeword number

$\zeta(i, j; \delta t) = \Sigma W(i,t) +W(i,t+\delta t)$ (4)

where δt determines the time balance, W(i,t) is the casing Ft's visual word likelihood to bunch i. Figure 4.3 demonstrates the correlogram development.

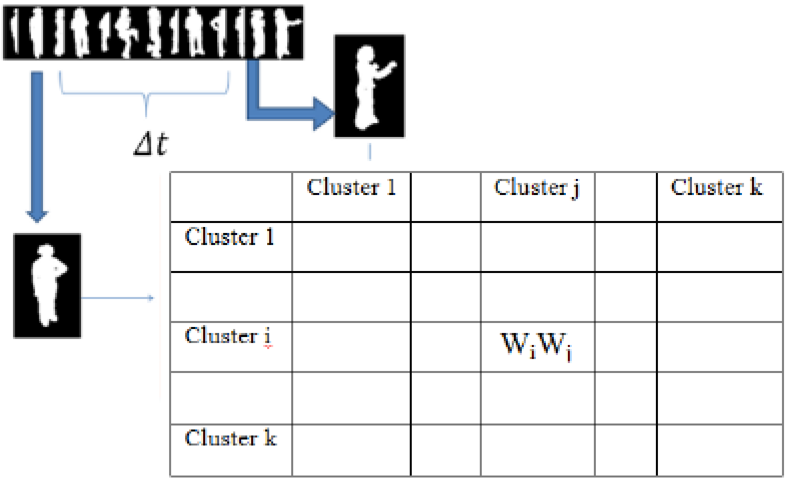


Fig 4.3 Construction of Correlogram

B. Dimensionality lessening

In this case a near investigation on bag of related technique over three distinctive feature vectors. Out of these three features, a nonlinear and two linear learning strategy is utilized. The first employments linear dimensionality lessening keeping in mind the end goal to discover the basic format of the information. Both of the techniques PCA and ICA are utilized for taking in an arrangement of Principal Components (PCs) for portrayal of the information. Now pre-handling of the outline vectors, the dimensionality diminishment procedures are continuing to the preparation database comprised of the outline vectors with a bulk measurement. In order to remove human action outline highlights, the most mainstream highlight extraction strategy connected in the videobased Human Action Recognition \ is Principal Component Analysis (PCA). It is the commonly utilized and surely understood of the standard multivariate *strategies. It comprises on a change from a space of high measurement to another with additional decreased measurement. On the off chance that the information sources are profoundly associated, there is excess data. PCA diminishes the measure of excess data by decorrelating the info vectors. The related info vectors, with high measurement, can be spoken to in a lower measurement space and decorrelated. It is a linear projection technique to decrease the quantity of parameters. It exchanges an arrangement of relate factors into anotherset of uncorrelated variables.*

Independent Component Analysis (ICA) is a computational strategy for isolating a multivariate signal into added substance subcomponents assuming the mutual statistical independence of the non-Gaussian source signals. ICA can be characterize by "Minimization of Mutual Information and Augmentation of non-Gaussianity" However dimensionality lessening procedures like PCA and ICA accept that the information basically lies on a linear manifold. Also, it is impossible that they lie on a straightforward straight complex. The principle issue utilizing PCA and ICA is that straight PCs can't speak to the non-direct nature of human movement.

The second strategy utilizes a non-linear dimensionality lessening system. In particular, spatio-temporal Isomap is connected to reveal the characteristic non-linear geometry of the information, and it is caught through figuring the geodesic complex separations between all sets of information focuses.

An essence of Isomap calculation is to find a proficient method to process the genuine geodesic separation between perceptions, provided just its Euclidean separations in the higher dimensional perception space. The thought is that Euclidean separation is around equivalent to the geodesic separation for near to focuses. For focuses which are *far away the geodesic separate must be processed by a progression of bounces. The Isomap calculation as proposed in [20] comprises of three principle steps.*

(1) Create the area chart G taken as a whole perception focuses.

(2) Calculate most brief ways in the chart involving utilizing either the Floyd's or the Dijkstra's calculation.

(3) Application of Multi Dimensional Scaling over the subsequent geodesic separation network for discovery of a d-dimensional inserting. When we apply PCA, there is an attempt to save covariance. As in this case we attempt to do perform the similar thing nonlinearly and the deployed approach is for making an attempt for protection of entomb point remove on the complex. Isomap and Principal Component Analysis perform it in an iterative form, having a go at expanding values for d and processing some kind of lingering fluctuation. Plotting this leftover against d will enable us to discover an infection point that shows a decent incentive for d.

Chapter 4

EXPERIMENTAL RESULTS

In this chapter, we bring the three standard and public datasets which are deployed throughout the work to validate the mentioned method.

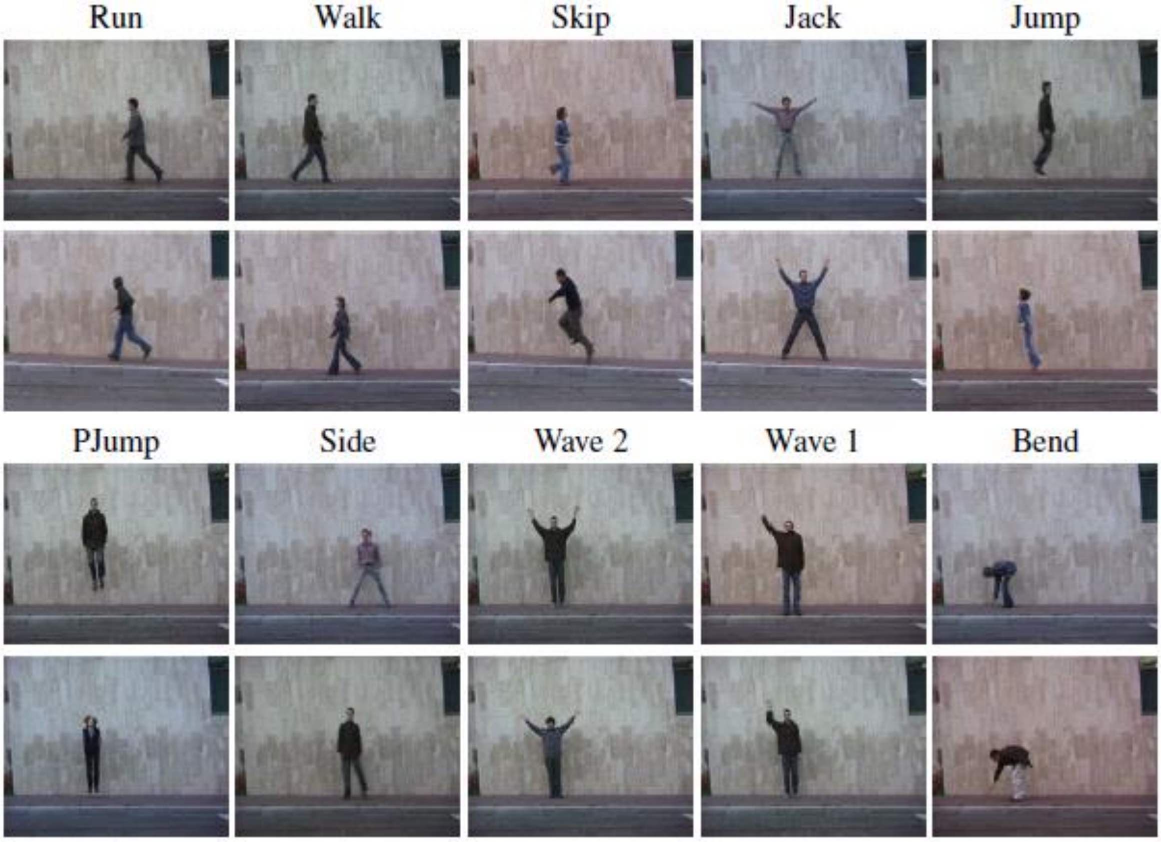


Fig 4.1. Sample frames of Weizmann dataset

The Weizmann Action Recognition dataset (to put it plainly, the Weizmann dataset) has been presented by Blank et al. [Blank 2005]5. It contains recordings of 10 sorts of human activities. The full rundown of activities is: run, walk, skip, hopping jack (instantly "jack"), bounce forward-ontwo- legs (presently "hop"), bounce set up on-two-legs (in a matter of seconds "pjump"), run sideways (presently "side"), wave-two-hands (in a matter of seconds "wave2"), wave-one-hand (in the blink of an eye "wave1"), *what's more, twist. Each activity is performed by 9 individuals. Recordings are recorded with 180 × 144 pixels spatial determination and 50 outlines for each second casing rate. Altogether, the dataset contains 90 video groupings.*

Id	Action Name	STIPs			
		min	max	avg	std
1	Run	72	302	172	91
2	Walk	100	512	311	154
3	Skip	72	424	204	129
4	Jack	87	486	221	140
5	Jump	72	395	129	102
6	Pjump	72	197	116	41
7	Side	83	176	133	32
8	Wave 2	72	72	72	0
9	Wave 1	72	72	72	0
10	Bend	72	96	76	9
	All Actions	72	512	153	115

Table 4.1 Weizmann dataset: Statistical (minimum, maximum, average and standard deviation) for the extracted features i.e STIPs

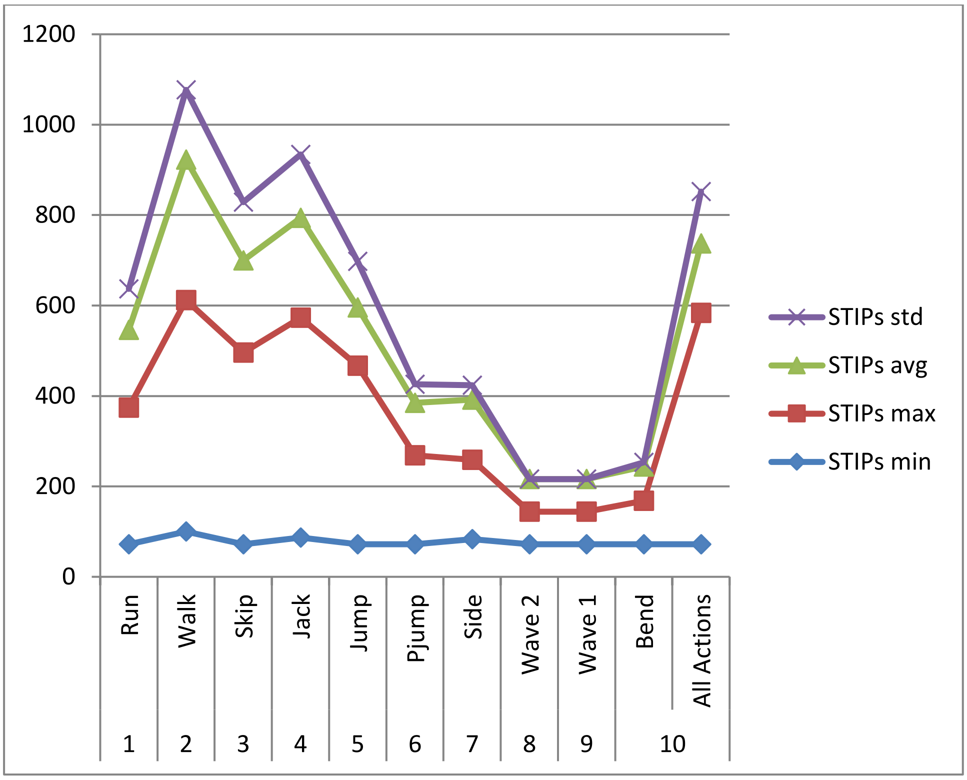


Fig 4.2. Weizmann dataset: Graph (minimum, maximum, average and standard deviation) for the extracted features i.e STIPs

The primary difficulties of the Weizmann dataset are: low determination recordings, different individuals, what's more, fabric varieties. Test video outlines from the Weizmann dataset are exhibited in Fig 4.1. We have likewise separated Spatio-Temporal Interest Points from this dataset, and we display the factual information about the removed components in Table 4.1. By and large, to assess an approach on this dataset, we utilize the forget one-individual cross-approval assessment plot, and we report comes about utilizing the mean class precision metric.

KTH dataset

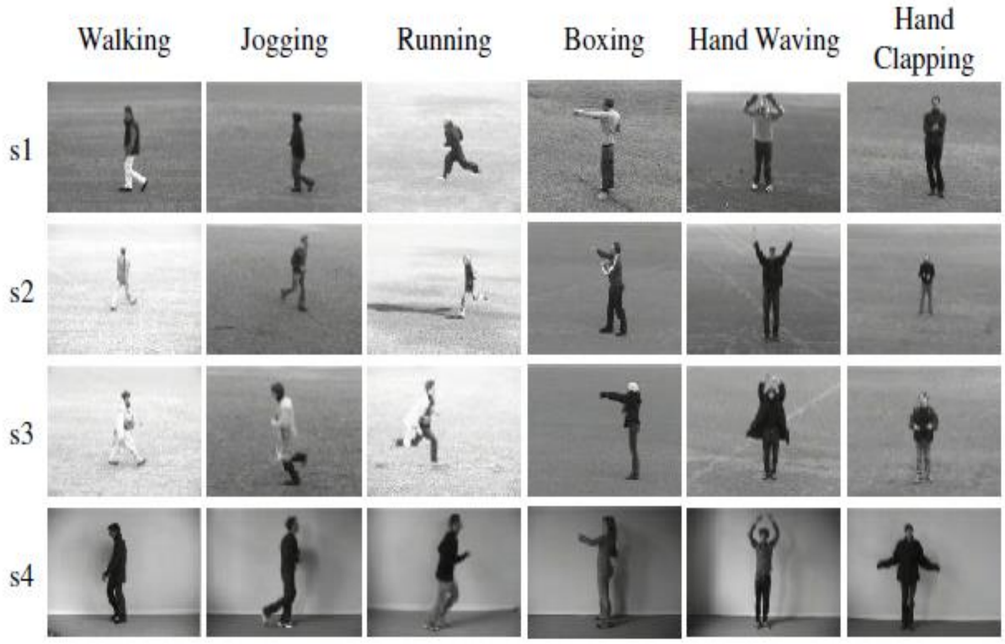


Fig 4.3 Sample frames of KTH dataset

The KTH Action dataset (to put it plainly, the KTH dataset) has been presented by Schuldt et al. [Schuldt 2004]6. It contains recordings of 6 sorts of human activities. The full rundown of activities is: strolling, running, running, boxing, hand waving, and hand applauding. Each activity is played out a few times by 25 distinct subjects. Each subject performs activities in four unique situations: outside (s1), outside with scale variety (s2), outside with various garments (s3), and inside (s4). All recordings are recorded over homogeneous foundations and are down-inspected by the creators to the spatial determination of 160 × 120 pixels. The successions are recorder utilizing a static camera with 25 outlines for every second edge rate, and have a length of four seconds overall. Altogether, the dataset contains 599 video documents.

ID	Action	STIPs

	Name	min	max	avg	std
1	Boxing	202	2087	800	402
2	Waving	292	2073	959	347
3	Clapping	231	1380	667	239
4	Walking	382	1590	865	223
5	Jogging	386	1972	793	264
6	Running	309	1910	662	262
	All Actions	202	2087	791	313

Table 4.2 KTH dataset: Statistical (minimum, maximum, average and standard deviation for the extracted features i.e STIPs

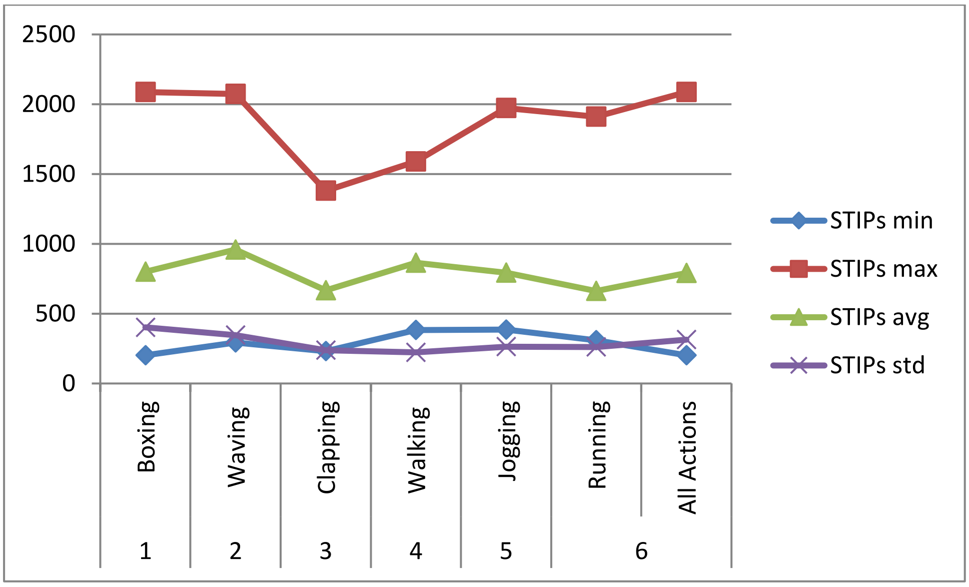
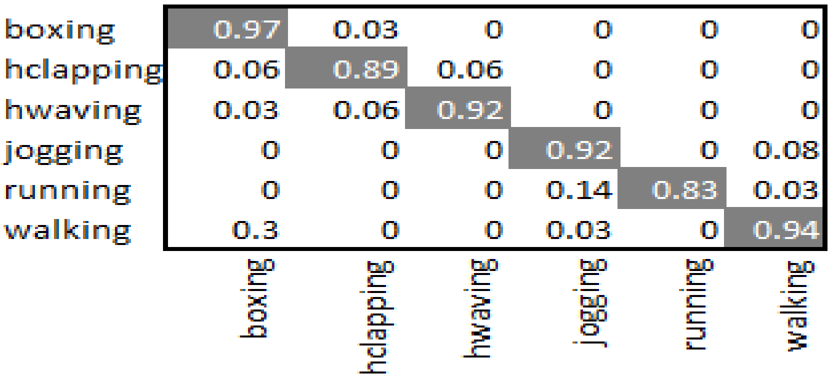


Fig 4.4 KTH dataset: Statistical (minimum, maximum, average and standard deviation for the extracted features i.e STIPs

The primary difficulties of the KTH dataset are: low determination recordings, scale changes, enlightenment varieties, shadows, distinctive individuals, diverse situations, material varieties, entomb and intra activity class speed varieties.



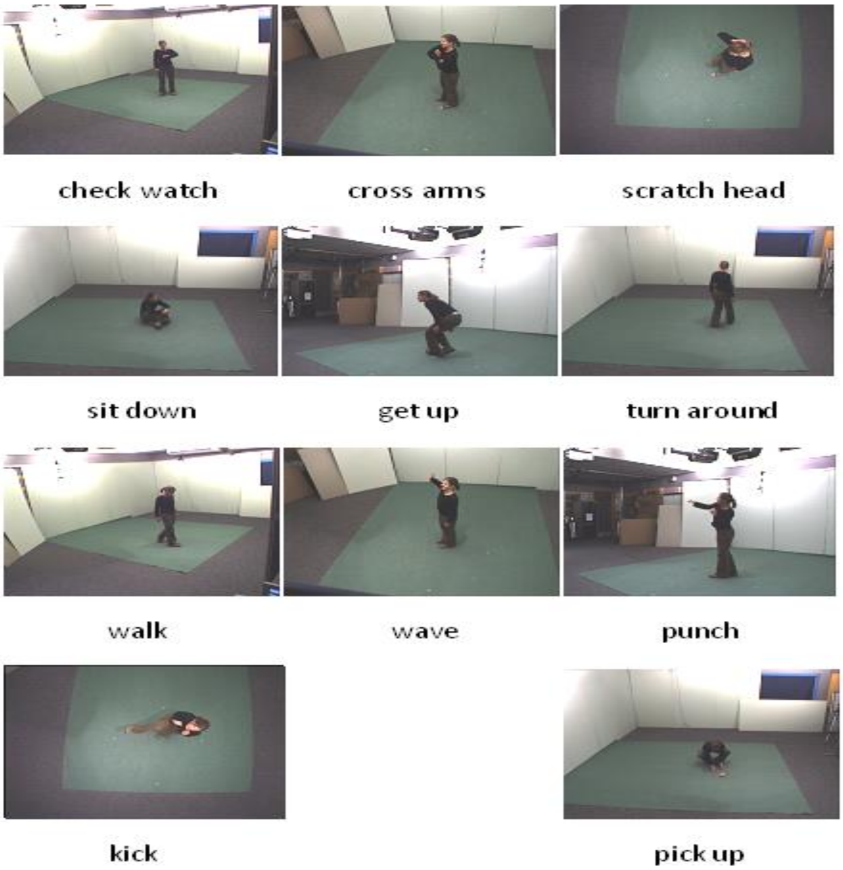
	boxing	hclapping	hwaving	jogging	running	walking
boxing	0.97	0.03	0	0	0	0
hclapping	0.06	0.89	0.06	0	0	0
hwaving	0.03	0.06	0.92	0	0	0
jogging	0	0	0	0.92	0	0.08
running	0	0	0	0.14	0.83	0.03
walking	0.3	0	0	0.03	0	0.94

Fig 4.5 Confusion matrix for KTH dataset

Test video outlines from the KTH dataset are exhibited in Figure 4.3. We have likewise separated Spatio-Temporal Interest Points from this dataset, and we exhibit the factual information about the separated elements.

IXMAS dataset

The INRIA IXMAS dataset [23] is a testing multi-see dataset for activity acknowledgment which is openly available. It contains 14 day by day live activities, each performed three times by 12 performers. Keeping in mind the end goal to test see invariance, the on-screen characters unreservedly change introductions in every execution with no data gave other than the names. Case pictures of 11 activities are appeared in figure 4.6. To be practically identical, in our trials, we have utilized a similar 11 actions and 10 subjects which are utilized as a part of [23] and [24].



check watch

cross arms

scratch head

sit down

get up

turn around

walk

wave

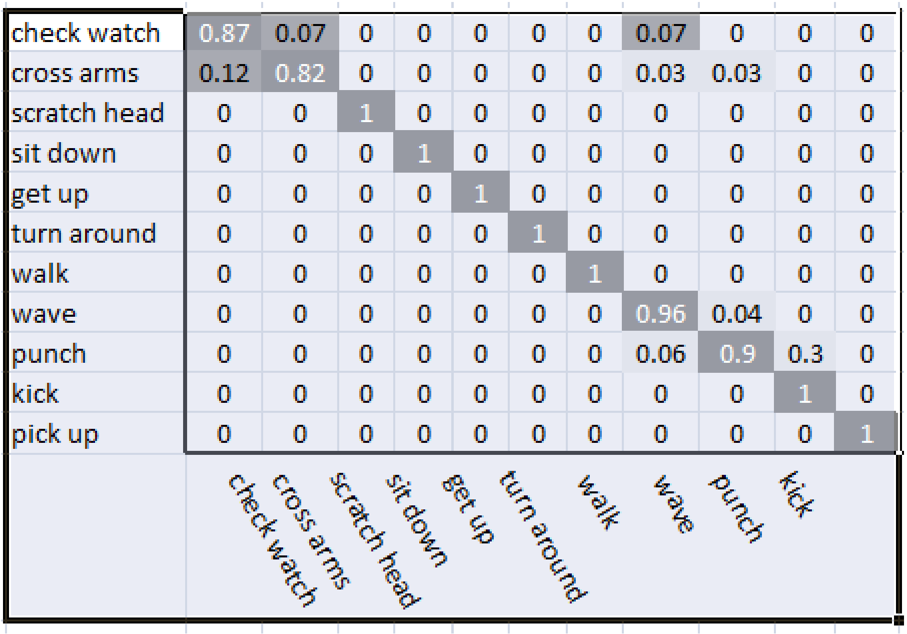
punch

kick

pick up

Fig. 4.6 Sample frames of IXMAS datasets

The movements in IXMAS dataset are caught utilizing five fire-wire cameras. Figure 5.4 shows illustration sees from the five cameras for kick activity. Like [23] and [24] we utilize forget one-individual cross approval. So in each cycle we prepare with the information of nine people and test with the information of the rest of the individual. This technique is rehashed for all the 10 people. The precision announced is the normal of all the 10 runs.



	check watch	cross arms	scratch head	sit down	get up	turn around	walk	wave	punch	kick	
check watch	0.87	0.07	0	0	0	0	0	0.07	0	0	0
cross arms	0.12	0.82	0	0	0	0	0	0.03	0.03	0	0
scratch head	0	0	1	0	0	0	0	0	0	0	0
sit down	0	0	0	1	0	0	0	0	0	0	0
get up	0	0	0	0	1	0	0	0	0	0	0
turn around	0	0	0	0	0	1	0	0	0	0	0
walk	0	0	0	0	0	0	1	0	0	0	0
wave	0	0	0	0	0	0	0	0.96	0.04	0	0
punch	0	0	0	0	0	0	0	0.06	0.9	0.3	0
kick	0	0	0	0	0	0	0	0	0	1	0
pick up	0	0	0	0	0	0	0	0	0	0	1

Fig. 4.7 Confusion matrix for IXMAS dataset

Chapter 5

CONCLUSION

Human activity acknowledgment in PC vision has pulled in solid research enthusiasm for late years in view of its promising applications. Computerized identification of a man's nuclear developments is called human activity acknowledgment. Various strategies had been introduced to distinguish the human activity because of the elements in basic portrayal of individual body. It gives a look greater than the huge investigates led around there and a complete study of the diverse scientific classifications recommended by various analysts. A framework is proposed for proficient activity acknowledgment in view of the idea ‗bag of associated postures' by assessing the connection between's consecutive stances in an activity on both direct and non straight element vectors. It is clear that human movement examination assumes a pivotal part in propelling PC vision. This venture has finished up with a basic however successful strategy for programmed individual acknowledgment from body outline with an augmentation of Correlogram based methodologies called Bag of corresponded groups (BoCP). It is a transiently neighborhood highlight descriptor. One of a kind method for considering worldly basic connections between's back to back human postures encoded more data than the conventional sack of features. display. Here, the framework demonstrated promising execution along with delivered preferred outcomes when utilizing ISOMAP over utilizing as it were low level elements and basic dimensionality lessening techniques like PCA and ICA. Despite the fact that ICA is demonstrated to be a decent technique for activity acknowledgment our test demonstrates that ICA is not well-suited for development of correlogram. For characterization when precision is *alarmed ISOMAP is ideal be that as it may, PCA could be utilized when speed is apprehensive. by means of additional modern component descriptors along with progressed dimensionality diminishment techniques, improved execution can be accomplished.*

References

[1] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, \Machine recognition of human activities: A survey," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, pp. 1473{1488, 2008

[2] R. Poppe, \A survey on vision-based human action recognition," Image and Vision Computing, vol. 28, no. 6, pp. 976{990, 2010.
[

[3] H. Ketabdar, M. Roshandel, and K. A. Y• uksel, \Towards using embedded magnetic _eld sensor for around mobile device 3d interaction," in International conference on Human computer interaction with mobile devices and services, 2010.

[4] J. Lee, S. Hudson, and P. Dietz, \Hybrid infrared and visible light projection for location tracking," in annual ACM symposium on User interface software and technology, 2007.

[5] P. Scovanner, S. Ali, and M. Shah, \A 3-dimensional sift descriptor and its application
to action recognition," in ACM International Conference on Multimedia,
2007.

[6] C. Cedras and M. Shah, \Motion-based recognition: A survey," Image and Vision Computing, vol. 13, pp. 129{155, 1995.

[7] M. Hoai, Z. Lan, and F. De la Torre, \Joint segmentation and classi_cation of human actions in video," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[8] Y. Zhong and M. Stevens, \Action recognition in spatiotemporal volume," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[9] J. Liu and M. Shah, \Learning human action via information maximization," in IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[10] Aaron F. Bobick and James W. Davis ―The Recognition of Human Movement Using Temporal Templates‖, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 23,NO. 3, MARCH 2001

[12] Lena Gorelick , Shechtman , Irani and Basri ―Actions as Space-Time Shapes‖ IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 29, NO. 12, DECEMBER 2007

[13] Xinghua Sun and Mingyu Chen Alexander Hauptmann ―Action Recognition via Local Descriptors and Holistic Features‖, Computer Vision and Pattern Recognition (CVPR) Workshops, 2009. CVPR Workshops2009. IEEE Computer Society Conference

[14] Di Wu, ―Silhouette Analysis-Based Action Recognition Via Exploiting Human Poses‖ IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 23, NO. 2, FEBRUARY 2013

[15] Fan Zhu, Ling Shao and Mingxiu Lin ―Multi-view action recognition using local similarity random forests and sensor fusion‖, ELSEVIER Pattern Recognition Letters 34 (2013) 20–24

[16] Ling Shao, Ling Ji, Yan Liu, Jianguo Zhang, ―Human actionsegmentation and recognition via motion and shape analysis‖, ELSEVIER,Pattern Recognition Letters 33 (2012) 438–445

[17] M Xiantong Zhen, Ling Shao ―Embedding Motion and Structure Features for Action Recognition‖, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY ,2013, Unpublished

[18] Sen Wang, Yi Yang, Zhigang Ma, Xue Li, Chaoyi Pang, Alexander G. Hauptmann ―Action Recognition by Exploring Data Distribution and Feature Correlation‖, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference

[19] Michael B. Holte, Bhaskar Chakraborty, Jordi Gonzàlez, and Thomas B. Moeslund,―A Local 3-D Motion Descriptor for Multi-View Human Action Recognition from 4-D Spatio-Temporal Interest Points‖ IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 6, NO. 5, SEPTEMBER 2012

[20] D. Weinland, R. Ronfard, and E. Boyer, ―Free viewpoint action recognition using motion history volumes,‖ Comput. Vision ImageUnderstand.,vol. 104, nos. 2–3, pp. 249–257, 2006.

[21] http:// isomap.stanford.edu/

[22]Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani and Ronen Basri. Actions as space-time shapes. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 2, pages 1395–1402. IEEE, 2005. (Cited on pages 22, 42 and 66.)

[23] Christian Schuldt, Ivan Laptev and Barbara Caputo. Recognizing Human Actions: A Local SVM Approach. In 17th International Conference on Pattern Recognition (ICPR), ICPR '04, pages 32–36, Washington, DC, USA, 2004. IEEE Computer Society. (Cited on pages 36 and 68.)

[24] D. Weinland, R. Ronfard, and E. Boyer, \Free viewpoint action recognition using motion history volumes," Computer Vision and Image Understanding, vol. 104, pp. 249{257, 2006.

[25] B. Peng and G. Qian, \Online gesture spotting from visual hull data," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 6, pp. 1175{1188, 2011.

[26] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in Proc. 2nd Joint IEEE Int.Workshop Vis. Surveillance Performance Eval. Tracking Surveillance ,Oct. 2005, pp. 65–72.

[27] I. Laptev and T. Lindeberg, "Space-time interest points," in Proc. ICCV,2003, pp. 432–439.

[28] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in Proc.IEEE Conf. CVPR, Jun. 2009, pp. 2929–2936.

[29] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in Proc. IEEE Conf. CVPR, Jun. 2008, pp. 1–8.

[30] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in Proc. BMVC, 2009, pp. 124.1–124.11.

[31] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in Proc. ECCV, 2008, pp. 650–663.

[32] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in Proc. IEEE Comput. Soc. Conf. CVPR, vol. 2. Jun. 2005, pp. 524–531.

[33] T. Goodhart, P. Yan, and M. Shah, "Action recognition using spatiotemporal regularity based features," in Proc. IEEE ICASSP, Mar.–Apr. 2008, pp. 745–748.

[34] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in Proc. British Mach. Vision Conf., 2008, pp. 995—1004.

[35] M. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," Carnegie Mellon Univ., Tech. Rep. CMU-CS-09-161, 2009.

[36] X. Cao, B. Ning, P. Yan, and X. Li, "Selecting key poses on manifold for pairwise action recognition," IEEE Trans. Indust. Inform., vol. 8,no. 1, pp. 168–177, Feb. 2012.

[37] J. Davis and A. Bobick, "The representation and recognition of human movement using temporal templates," in Proc. IEEE Comput. Soc. Conf.Comput. Vision Patt. Recog., Jun. 1997, pp. 928–934.

[38] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Proc. 10th IEEE ICCV, vol. 2. Oct. 2005, pp. 1395–1402.

[39] L. Wang and C. Leckie, "Encoding actions via the quantized vocabulary of averaged silhouettes," in Proc. Int. Conf. Patt. Recog., 2010, pp. 3657–3660.

[40] L. Shao and X. Chen, "Histogram of body poses and spectral regression discriminant analysis for human action categorization," in Proc. BMVC, 2010, pp. 88.1–88.11.

[41] H. Qu, L. Wang, and C. Leckie, "Action recognition using space-time shape difference images," in Proc. 20th ICPR, 2010, pp. 3661–3664.

[42] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in Proc. IEEE Comput. Soc. Conf. CVPR Workshops, Jun. 2009, pp. 58–65.

[43] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in Proc. WorkshopStatist. Learn. Comput. Vision (ECCV), 2004, pp. 17–32.

[44] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek, "Visual word ambiguity," IEEE Trans. Patt. Anal. Mach. Intell., vol. 32, no. 7,pp. 1271–1283, Jul. 2010.

[45] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in Proc. IEEE Conf. CVPR, Jun. 2010, pp. 2559–2566.

[46] X. Zhou, X. Zhuang, H. Tang, M. Hasegawa-Johnson, and T. Huang,"Novel Gaussianized vector representation for improved natural scenecategorization," Pattern Recog. Lett., vol. 31, no. 8, pp. 702–708, 2010.

[47] L. Shao, D. Wu, and X. Chen, "Action recognition using correlogram of body poses and spectral regression," in Proc. Int. Conf. Image Process., 2011, pp. 209–212.

[48] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih, "Image indexing using color correlograms," in Proc. IEEE Comput. Soc. Conf. Comput. Vision Patt. Recog., Jun. 1997, pp. 762–768.

[49] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in Proc. 17th ICPR, vol. 3. 2004, pp. 32–36.

[50] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," Comput. Vision Image Understand., vol. 104, nos. 2–3, pp. 249–257, 2006.

[51] D. Weinland, M. O¨ zuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in Proc. ECCV, 2010, pp. 635–648.

[52] M. Varma and B. Babu, "More generality in efficient multiple kernel learning," in Proc. 26th Annu. Int. Conf. Mach. Learn., 2009, pp. 1065–1072.

[53] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in Proc. IEEE Conf. CVPR, Jun. 2011, pp. 489–496.

[54] I. Junejo, E. Dexter, I. Laptev, and P. P´erez, "Cross-view action recognition from temporal self-similarities," in Proc. ECCV, 2008, pp. 293–3