Title Feature Based Dynamic Intra-video Indexing

Name Muhammad Nabeel Asghar

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.

3403962805

Feature Based Dynamic Intra-video Indexing

by

Muhammad Nabeel Asghar

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Institute for Research in Applicable Computing (IRAC)
Department of Computer Science & Technology
University of Bedfordshire

September 2014

3403962805
006 . ASG

Abstract

With the advent of digital imagery and its wide spread application in all vistas of life, it has become an important component in the world of communication. Video content ranging from broadcast news, sports, personal videos, surveillance, movies and entertainment and similar domains is increasing exponentially in quantity and it is becoming a challenge to retrieve content of interest from the corpora. This has led to an increased interest amongst the researchers to investigate concepts of video structure analysis, feature extraction, content annotation, tagging, video indexing, querying and retrieval to fulfil the requirements. However, most of the previous work is confined within specific domain and constrained by the quality, processing and storage capabilities. This thesis presents a novel framework agglomerating the established approaches from feature extraction to browsing in one system of content based video retrieval. The proposed framework significantly fills the gap identified while satisfying the imposed constraints of processing, storage, quality and retrieval times. The output entails a framework, methodology and prototype application to allow the user to efficiently and effectively retrieved content of interest such as age, gender and activity by specifying the relevant query. Experiments have shown plausible results with an average precision and recall of 0.91 and 0.92 respectively for face detection using Haar wavelets based approach. Precision of age ranges from 0.82 to 0.91 and recall from 0.78 to 0.84. The recognition of gender gives better precision with males (0.89) compared to females while recall gives a higher value with females (0.92). Activity of the subject has been detected using Hough transform and classified using Hidden Markov Model. A comprehensive dataset to support similar studies has also been developed as part of the research process. A Graphical User Interface (GUI) providing a friendly and intuitive interface has been integrated into the developed system to facilitate the retrieval process. The comparison results

of the intraclass correlation coefficient (ICC) shows that the performance of the system closely resembles with that of the human annotator. The performance has been optimised for time and error rate.

Acknowledgement

All praise to Almighty Allah, who is enormously benevolent and humble to all human beings. My gratitude is accredited to Almighty Allah.

I want to dedicate this thesis to sweet loving memory of my father in heaven and dearest mother for their love and encouragement, they are the true motivation of my life, who offer me bounteous of opportunities to enable me to learn extensively and purposefully. I am thankful to my whole family especially my siblings, maternal and paternal uncles, aunts, cousins and my in-laws. But most importantly, I am grateful to my loving wife who has always supported and helped me and stood besides me through thick and thin.

Here, I owe a debt of my gratitude to my supervisors Dr. Fiaz Hussain and Rob Manton for their valued criticisms, suggestions, constant encouragement and great devotion of their time to discuss and develop this work with me. I also owe special appreciation to Director of Institute, Prof. Edmond C. Prakash for his valuable guidance throughout these years.

Special gratitude is kept for Bahauddin Zakariya University Multan,Paksitan who generously awarded scholarship which permitted me to concentrate fully on my research from 2010 to 2013.

In the end, I want to mention about my dearest friends Usman Ghani, Adeel Nawab, Munam Ali Shah, Adeel Khan, Sufyan Niazi, Faisal Qureshi, Kamran Abbasi, Adnan Qureshi and the list goes on. They were always with me whenever I need them.

Thanks a lot to all of you for being the inspiration, motivation and the beacon of hope for me. Thanks to all the people who have helped me with many things.

Declaration

It is hereby certified that the substance of this thesis entitled Feature Based Dynamic Intra-video Indexing is the original work of the author and due acknowledgements and references have been made, where necessary, to the work of others. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

I also certify that the material in this thesis has not been already accepted or being currently submitted in candidature of any degree at any other University.

Signed:

Full Name: Muhammad Nabeel Asghar

Date : September 2014

Abbreviations

AAM	Active Appearance Model
AFLW	Annotated Facial Landmarks in the Wild
AV	AudioVisual
CC	Closed Caption
CECH	Colour Edge Co-occurrence Histogram
CRG	Colour Ratio Gradient
CWT	Continuous Wavelet Transform
DCT	Discrete Cosine Transforms
DDL	Description Definition Language
DS	Description Scheme
DWT	Discrete Wavelet Transform
EHD	Edge Histogram Descriptor
EM	Expectation Maximization
FHD	Fine Home Displays
FIR	Finite Impulse Response
FN	False Negative

FP	False Positive
GMM	Gaussian Mixture Model
GT	Ground Truth
GUI	Graphical User Interface
GWHI	Gaussian Weighted Histogram Intersection
HCI	Humancomputer interaction
HD	High Definition
HI	Histogram Intersection
HMM	Hidden Markov Models
HSV	Hue Saturation Value
HVC	Hue Value Chroma
HVGA	Half Video Graphics Array
ICA	Independent Component Analysis
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
kNN	k-Nearest Neighbours
LBG	Linde-Buzo-Gray
LDA	Linear Discriminant Analysis
MEI	Motion Energy Image
MHAD	Berkeley Multi-modal Human action Database
MHI	Motion History Image
MLD	Moving Light Display

MPEG	Moving Picture Experts Group
MPHM	Merged Palette Histogram Matching
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
PPV	Positive Predicted Value
QCIF	Quarter Common Intermediate Format
QVGA	Quarter Video Graphics Array
RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
RR	Relevance Ranking
SDLC	Software Development Life Cycle
SIFT	Scale Invariant Feature Transform
SKIG	Sheffield Kinect Gesture
SQL	Structured Query Language
SVGA	Super Video Graphics Array
SVM	Support Vector Machine
SXGA	Super eXtended Graphics Array
TN	True Negative
TP	True Positive
TPR	True Positive Ratio
TREC	Text REtrieval Conference

TRECVid	TREC Video Retrieval Evaluation
TV	Television
UVGA	Ultra Video Graphics Array
UXGA	Ultra eXtended Graphics Array
VCAGA	Video Corpus for Action Gender Age
VGA	Video Graphics Array
VI	Video Indexing
VO	Video Object
VOL	Video Object Layer
VOP	Video Object Plane
VQ	Vector Quantization
WSXGA	Wide Super-eXtended Graphics Array
WXGA	Wide eXtended Graphics Array
XGA	eXtended Graphics Array
XML	Extensible Markup Language

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The exponential increase in the production of digital images and video content during the last few decades result in the availability of data in the form of movies, photo albums and broadcast news videos. Video material may consist of feature films, hand-held movies or broadcast news and audio and visual scenes. These modalities can be used for discovering the underlying semantics of videos. According to the Cisco estimation, more than 90% of consumer traffic will be based on video content by 2015 [2].

Any video processing paradigm must incorporate the strengths and weaknesses of the media by which it is conveyed. Different media sources have specific attributes and structure; and have their own advantages and disadvantages. For instance, hand-held video camera movies are mostly family oriented, e.g, feature films are often for entertainment, lesson or recreation providers, while historical videos are informative and television broadcasts are better at tracking and developing news stories. Each medium has its own pros and cons. In general personal videos are unstructured; on the other hand broadcast videos are highly structured.

Video data comprises of visual content and audio stream and closed caption information in textual form which results in huge multidimensional information. This

multidimensionally urges for qualitative filtering of data to extract the required and relevant information as desired by the user. The amount of time required for processing and extracting the relevant information from huge and complex data is also eye provoking. There is a need for time limitation to spend on extracting the required and relevant information. This distillation process needs a comprehensive understanding of visual processing paradigm to dig the required information from huge visual data. Tags in the textual form such as a title, keywords, a brief summary or any related description often provide good information.

Humans can easily understand a video as they develop natural capabilities of understanding a visual scene based on its contents and then perceiving it using their domain knowledge, but computers are far behind in this capability. Recently, studies have been done related to object recognition and categorization. Computers can identify objects and estimateactivities in video up to a certain accuracy but currently there is a need for research inautomatic understanding of visual contents and visual scene indexing . Video data should be searched with same efficiency as text; only storing is not enough. Finding a specific video from a large repository is a difficult and frustrating task, there is need of retrieving and searching techniques even query results are inconsistent. Also there is a requirement for searching specific part of video rather than navigating through the whole video. All these factors arise the need of proposing technique for classifying and retrieval of video content which will efficiently give better result in video indexing systems.

1.1 Aim and Objectives

The aim of the study is to design, implement and evaluate feature extraction, video indexing and retrieval approaches focussing on content based feature extraction using low-level feature and high level semantics in order to minimise the errors and

maximize the accuracy of the indexing system without requiring human intervention. The aim leads to the following objectives:

1. Performance evaluation of feature based video indexing techniques.
2. Performance evaluation of feature extraction techniques for indexing.
3. Developing an automatic search tool to facilitate browsing and retrieval.
4. Designing and developing a framework to streamline the extraction and population of a database during training and retrieval during testing.
5. Development and evaluation of the video dataset.

The interest is in a methodology and a proof of concept rather than developing a commercial application for video indexing. Development of compact indexing and retrieval tool which return all desired locations in a vast amount of visual data. It could be used in the movie post-production industry, personal photo and video collections or Internet image and video archives.

1.2 Contribution

The key contributions of this study are:

1. The performance improvement of the feature extraction techniques for video indexing and retrieval.
2. The automatic indexing approach with no human intervention during extraction and indexing.

3. The framework providing the integration of feature extraction, video indexing and retrieval for improved user experience for content based video retrieval.

4. The creation of the dataset for evaluation for content based video retrieval research mainly focussing on age, gender and activity.

5. Design and development of content based video indexing and retrieval system.

1.3 Thesis Outline

The thesis is structured as follows.

Chapter 2 covers the literature survey specially in the context of video indexing and retrieval. The focus remained on the overall process involved in video indexing and retrieval specifically video segmentation, video retrieval, feature extraction, video annotation, video queries and classification of videos by means of certain queries.

Chapter 3 consists of the framework, where the holistic architecture of the video indexing and retrieval system has been illustrated and defined components and structure wise.

In chapter 4, the video corpus generation process is discussed phase wise and covers the overall process followed.

In chapter 5, the shot boundary detection technique is discussed and the procedure has been devised to make use of it for upcoming steps.

Chapter 6 covers the feature detection and analysis process for human identification, age recognition, gender recognition and activity recognition. Furthermore, the learning, tagging and indexing process are discussed there.

Chapter 7 covers the user based evaluation of the overall system. There we discuss the basic issues in human based evaluation in context of video retrieval and provide the solution as the proof of concept.

The conclusion of the overall thesis is given in chapter 8 with insights drawn from the research for future work.

Chapter 2

Literature Review

The basic multimedia information is required for dynamic video indexing and retrieval. There are necessary components for storing, sorting and accessing multimedia contents. They also help in finding the desired components to form a multimedia repository with an ease [3]. Besides many multimedia resources, video is a key component which comprises mainly upon three major parts. The first one is that the vigorous video provides rich contents then that of images. Secondly there is huge amount of unprocessed data. Lastly, video structure is complex as it comprises of multi-modal information. Therefore, all these attributes make the retreival process complex and time consuming.In the past, retrieval and indexing had been controlled by the annotation of manual keyword as well as the database of the videos was very small. Although in the present era database is getting enormous, the content-based retrieval and indexing are needed with less human interaction to analyse videos automatically.

There are a lot of applications for video indexing and its retrieval. For instance, inspection of the pictorial automated commerce (user awareness tendencies investigation for its selections and collations), correspondences of examination in numeral institutions, instant browsing of video folders, summary occurrence investigation [4],

logical organization of videos of the web (beneficial video examine along with that locating detrimental videos), and audio-visual assessment which has motivated interests of the researchers. Two worth paying attention to research activities going on regularly are:

i) The National Institute of Standards and Technology (NIST) fund the biggest almanac video retrieval conference under the name of TRECVID. TREC(Text REtrieval Conference) has a sub-domain for storing video type of data known as Video Retrieval Evaluation (TRECVid). This is working since 2001 for the upgrading recommendation in audio-visual retrieval and analysis. Substantial amount of experimented video has been stored by TRECVid under this project. A lot of participants have given their algorithms of video retrieval on the basis of content of the work composition [5], [6], [7]

ii) For standard video, the main objective isto endorse suitability between representation of the interfaces for video fillings to improve and help to profligate improvement as well as flawless repossession algorithms of videos.

TV-Anytime Standard and moving picture experts group (MPEG) are the videos crucial standards [8]. At hand there have been lots of surveys using MPEG-7 for the video contents arrangement for the characteristics extraction or intended to the video objects explanations and compressed domain [9]. In a video there is possibility of visual channel along with the auditory channel. Consequently in videos the information is accessible [10], [11]:

a) In meta-data of the video it is patent as words embedded in period, performers, videos, heading, manufacturer, summary, along with that program length, video format, categorizer magnitude, and exclusive rights etc.

b) The acousticstatistics are given in the auditory channel.

c) The speech transcription is attainable with the help of speech recognition and caption, texts are scrutinized with the help of character recognition methods.

d) In the images it is a recognized phenomenon, that from visual channels the ocular data is gained and after it has been uploaded on web page, the page versions are linked with video. .

The architecture of video retrieval and indexing framework can be visualized in figure 2.1.

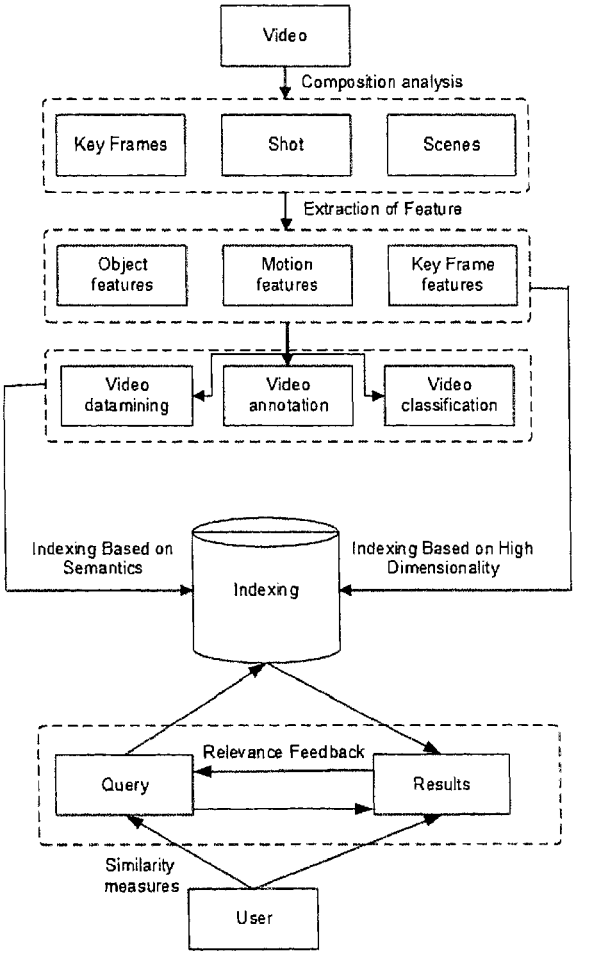


FIGURE 2.1: Video indexing and retrieval framework

The Main concepts included within the framework are as follows:

- Exploration of the structure: to discover the abstracts of the vital settings, shot margins, and portions of the scene
- The video segmented parts of units (scenes/stilled): in vital settings, features of motion, and feature of objects are comprised of static features
- Audio-visual explanations: For the manufacturing of a semantic index of video, extracted types and quarried information have been incorporated. Video arrangements deposited in the database have high range video manifestation of imminent vector, semantic and aggregate index;
- Query: To search needed videos the database of the video is used with the help of index and the video corresponding procedures;
- Visual browsing and response: In the formation of video review the searched videos are displayed in answer to the user query along with that the searched data have been enhanced to the relevant response.

2.1 Components of Video Retrieval Systems

Video structure is multi-modal and due to the complexity of the video data, a complete understanding of its quantities is required in order to manage it [12][13]. There are some distinct and complex qualities, which differ video from other types of data. Firstly, video has higher resolution, larger set and volume of the data that can be stored on disk, complex analysis, and requires more assimilation struggle due to the binary system of storing a video which is however alphanumeric in comparison to simple text-based data. Secondly, text can be categorised as only non-spatial static data and image is spatial static as compared to the video which has spatial and temporal context in nature. Furthermore, video semantic is not structured and have a very complex analogy.

Audiovisual (AV) presentation and semantic content are the vital ingredient of the video document. Message, knowledge, story or entrainment is the main semantic content which is delivered by the video data. Semantic part of the video data can either be implied explicitly or implicitly due to its complexity. Implicit semantic can be understood by the viewers implying their knowledge of seeing and hearing audiovisual presentation; however, explicit semantic can automatically be understood more easily. For example, text displayed to give information about the crew casts in a film-type video to viewers. Some acute audio and video information such as volume, colour, pitch, object motion, object relationship and texture can be extracted by the viewer from the AV tracks. Audiovisual presentation also includes some long textual displays to give information, which allows users to understand what is currently viewed by audiovisual presentation. For example, the event, place where the event took place and the issue under consideration along with who is involved in a shot of a video of the news to help viewers understand. Some closed caption (CC) tracks also displays sequential texts in an audiovisual presentation in a synchronising manner in some broadcast videos. Therefore CC track is also known as the subtitle or script of the spoken words.

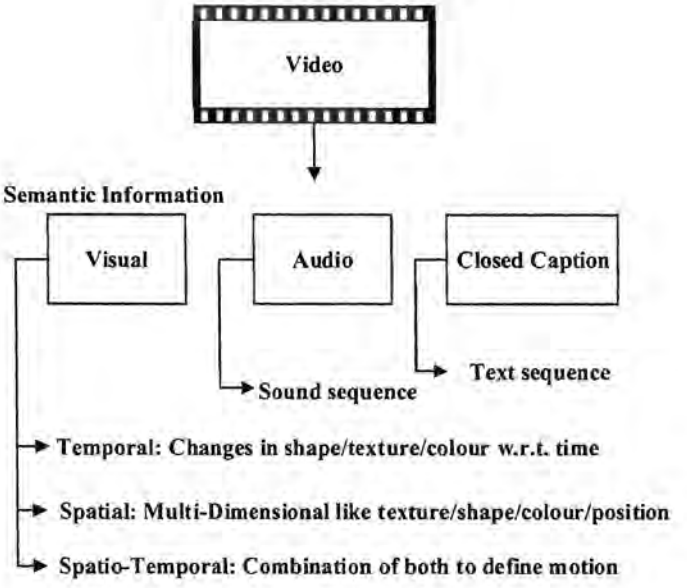


FIGURE 2.2: Video components

Figure 2.2 illustrates the top-level idea about the video contents data. Figure 2.3 shows the very simple example on delivering semantic contents based on spatio-temporal presentation within a video. Since shots are just displayed as an arrangement of frames in the figure 2.3, which shows that a car has travelled from right side to left side, which is the spatial-temporal information, whereas spatial-static information is depicted by a tree located in the background in middle of the frame.

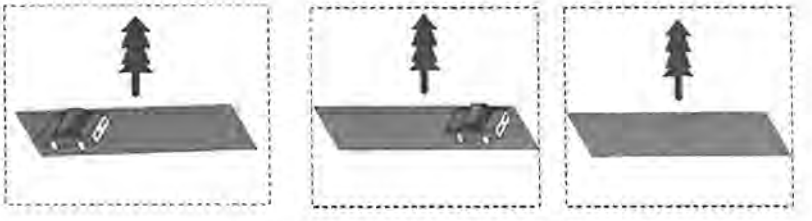


FIGURE 2.3: Spatio-temporal semantic

It is important to keep in mind that AV components of the video data are not always significant to deliver the semantic content. Therefore, it relies on the use and the

purpose of the video data. For example, the information on the motion and the place among the players are substantial important to understand the techniques and the strategy of the game in a soccer match. Some distinctive AV features can represent identical semantic contents, depending how the video was initially produced, on the other hand; due to the individual annotator, interchangeable AV feature can depict distinctive semantic contents. Consequently, when different channels are perceived, semantic information can be more nicely illustrated, same as the human perception of the video data.

The universal structure of video retrieval constituents can be explained as below. Retrieval and browsing can be figured out by user/application requirements. Effectiveness and completeness of the indexes are the key to successful retrieval. Indexing methods are decided by the information extracted through semi-automatic or automatic content distillation. Due to rich and multidimensional information included the video, in order to reach the efficient and compressed reproduction of the video data, it has to be summarised and sculpted. The structure of a video sequence has to be analyzed and separated into different layers such as shot and scene before the analysis of the video content. It is normally veryhard to elaborate each and every component differently because every single component is affected by the design of the component. Let's say, if retrieval is relied upon top-level semantic features such as particular sport highlight segments like the goal, the hierarchy of the key events will be based on indexing. Each event can be summarized by using the facial information of the participating actor during the event under consideration. However, event itself can be elaborated using some statistical measure, for example, the excitement ratio (the more obvious the event would be, if the excitement ratio were higher). Consequently, the content extraction process triggers ultimately to identify and classify the event(s) that is contained within each play break segment.

2.2 Video Structure Segmentation

Indexing video could be done at different density. If it's done on the whole video it will be raw and not fine-tuned and if it's done on every frame of the video then will be too hectic and long. So the researcher normally do indexing based on group of frame having some semantic co relation [14]. The two broad categorises for video segmentation includes border recognition of shot and scene change division.

2.2.1 Shot Boundary Detection

In video, shot which is the successive sequence of edgings, with camera they have been taken. In the beginning and the end the action is performed which defines the borders of the shot [15]. There have been plenty powerful gratified associations in the boundaries of the shot. Shots have been considered as, the main object for the establishment of the sequenced video content as well as the basics for the modern level annotation semantic and tasks of retrieval. Mostly, the categorization of the shots have been as cut in the shot evolution successive which is fast along with that the ongoing transition contains wipe, dissolve and fade light & high; it reaches above the amount of frames. In comparison to the detection of the cut the continuous detection transition has been very challenging. In prospect of the shot boundary detection and on shot boundary video detection there has been a transitory history of the study and reviews [16], [6]. For the shot boundary detection the fundamental groups methodologies have been presented analytically as well as with its evaluations, potentials and drawbacks. In the beginning, most of the times the video features have been pulled out from every frame for the detection of shot boundary, secondly the resemblances have been calculated among the frame with the help of abstracted features, and lastly those shot boundaries have been identified, which have no equivalency in frames. Additionally, three focal phases are

discussed for detection of the shot boundary: similarity size, abstraction characteristics, [17], and conclusions. To point out the detection boundary, the shot features have been used, for instance the vectors motion, histogram colour [18], histogram block colour and change edge ratio [19] [20],with the features, like , Scale Invariant Transform Feature (SIFT)[21], saliency map [22], corner points [23], etc. Minute camera motions between the histograms of colour have been very strong and sensitive as compared to large camera motion, though the similar acts of the shots are not distinguishable. The Edge features have radiance alterations and gesticulation that is why it has great constant in comparison to the histograms of colour, by the help of features motion the object inspiration and motion camera is controllable. The simple histograms colours are normally not outclassed by features of the edge although features motions are more complex structures [16]. For the measurement of the resemblance among frames the feature extraction is the second stage for the detection of the boundary shot. Euclidean distance, 1-norm cosine dissimilarity, chi-squared similarity and the histogram intersection, they have been the modern resemblance metrics for the vectors of the extracted features [24], [25], [26], alongside the space in the earth movers [18], the information given contains numerous new similarities measurements [27], [28], [29]. For measurement of the similitude between the frames successive pair wise resemblance are used as well as by window similarity measurement the frames in the windows parallels are measured [26]. For the contextual inculcation of the information on the bases of the swaying local noises decline the window based resemblance measurements have been used in disturbances. In comparison to the pair wise resemblance measurement the added calculation has been required. The identification of the shot boundary is defined in frames as by using the calculated resemblance. In the current progression, the shot boundary discovery is divisible in threshold based and statistical leaning.

2.2.1.1 Approach of Threshold

The pair wise measurement comparison for the resemblance of the frames with the predefined threshold, with the help of threshold based approach, the shot boundaries have been detected [30], [31]. When the similarity has become lower as compare to the threshold at that time the boundary is identifiable. There is the possibility of three categories of the threshold, universal, adaptive, the amalgamation of global and adaptive:

1. In the universal threshold based algorithms the comparable threshold has been used which is mostly fixed empirically on the video [31]. The main drawback of the threshold universal based algorithms is the rarer operational inculcation of the local content in the estimation of universal threshold this has influenced the precision of boundary detection.

2. The adaptive threshold based algorithms computes the threshold locally in a slithering window [22], [18], [29]. The presentation of the detection normally improves when an adaptive threshold is used as the substitution to the universal threshold [32]. Though in comparison to the adaptive estimation threshold the universal threshold estimation is simple. In adaptive threshold users must contain the information about the videos description for the parameters selection, for example, the size of the slithering window.

3. With the amalgamation of the adaptive and global algorithms it perfects the native thresholds with the help of the account value of universal threshold. With the recognition of the transition Cut, disband transition and gaudy detection have been demarcated by Qusenot *et al.*[33]. Both these global threshold roles which have been encountered from off trade exactness and remembrance.Functions value alters locally although the algorithm needs alteration for the dualistic global thresholds. The restrictions of the algorithms have

been that the practical dealings of locally adaptive thresholds and dual global thresholds have not been able to resolute it easily.

2.2.1.2 Statistical Learning-Based Approach:

By the approach of the arithmetical learning-based the detection of boundary shot has been considered as the taxonomy duty. The frames have been categorized on the source of its features; this is dependent on no shot or shot disparity. In this approach both directed (supervised) and non-directed (not-supervised) learning have been used.

a) Supervised classifiers of the learning based: Adaboost and Support Vector Machine (SVM) have been frequently used and organized or administered classifiers for the discovery of the boundary shot.

   1) SVM: SVM has been incorporated as dual class organizer to separate cuts from non-cuts by Chavez *et al.* [34]. A core function has been used to surmount the influence of the modification in entity's firm movement and illumination for the subversion of the assemblies in the high dimensional zone[24], [35]. In a sliding window by Zhao *et al.* [36] dual SVM classifier have been subjugated for the discovery of regular change and congruently cuts. The extraction was done primarily by Ling *et al.* [37] after that plenty features of the frames have used the SVM for the frame's cataloguing in three groups: measured transition, cut and others. Based classifier of SVM and the method based threshold has been combined Yuan *et al.* [16] and Liu *et al.* [28]. The precincts for the participants have been chosen mainly with the method of threshold based. After- wards for the boundaries identification, comprehensively SVM classifier has been used [38] for the discovery of the

shot boundary. Also, there has been a constant use of algorithm based on SVM. Its advantages are subsequently:

a) Information preparation and the good generalization conservation has been completely used by them

b) With the use of the core function the features of huge statistics have been handled effortlessly

c) Many SVM codes have been accessible previously.

2) Adaboost: The detection of cut has been proposed by Herout *et al.* [39]for detection task design in which the Adaboost algorithm has been used. Zhao and Cai [19] smeared it in the compacted dominion for the detection of the shot boundary. In the beginning the ambiguous classifier adaptation of the colour and pictures in motion endures approximately confidential. Afterward each frame has been defined as a cut, and classifier of the Adaboost has been incorporated for the dawdling, or nil alteration of the frame. The foremost benefit of the Adaboost boundary classifiers has been that in the vast amount the features are controllable: They are a part from the features chosen by these classifiers for the classification of boundary.

3) Others: The supervised algorithms are used for the detection of the shot boundary. For instance Cooper *et al.*[26] used dual (kNN) k Nearest-Neighbour segmentation in it the resemblances of frames have been used as involvement in the specific chronological intermission. Hidden Markov Models (HMM) have been functionalized by Boreczky and Wilcox [40]to discover the deliquesced, dwindled, prototypical shot cuts, zooms of the distinct conditions & pans. The previously declared profits are about the supervised learning method which describes that in the threshold based approaches thresholds setting is not required and for the detection accuracy upgrading diverse kinds of features are amalgamated. Their profound reliance shortcomings

on the appropriate assortment of preparation were fixed with the samples of both destructive and optimistic.

b) Unsupervised learning based algorithms: they have been categorized into the algorithms based on similarity of frame and simply base of the boundary. The similarity measurements have been gathered because of the similarity based algorithms of the frame it is divided in two clusters between the pair of the frames: The collection of the resemblances has inferior standards as it counterparts with the superior correspondence values of the shot boundaries that counterpart the non-boundaries [41]. K-means and vague K-means have been engaged in algorithms clustering's. Every shot has been preserved by algorithms of the frame grounded as the collection of surrounds which are comparable in the content of the chromatic. To the assemblage of different corresponding shots of frames by Chang *et al.* the clustering groups have been used [21]. K-means clustering used by Lu *et al.* [25] and also has been used and phantom gathering for the cluster of frame to detect different shots used by Damnjanovic *et al.* [42]. In the worth of gathering based approaches the training assets have not been needed. The sequential development arrangement information has not been deposited. The limitation is that they have incompetency in recognizing diverse kinds of measured alteration. The taxonomies of the shot boundary detection approach are the uncompressed domain-based and compressed domains based. By avoiding the time deteriorating video decompression, the trampled domain features have become handy, its example is the cosine discrete coefficients transform are used [43], [44], [19].For the shot boundary detection, motion vectors in DC image and MB types are working. The compression standards are for the great dependence on the compressed domain approach base. They have been reduced in exactness in comparison to the uncompressed approach of the domain base. By the gradual

detection the extra responsiveness can be received. Ngo proposed multi resolution analysis based on dissolves [45]. In accordance to the gradual transitions, discrepancy delivery arc of information edge in sequences frame detected by Yoo *et al.* [46]

2.2.2 Scene Segmentation

Segmentation of scenes is called section component division as well. Typically the story segment/section which has been recognized as the constant shots which are lucid along with the specific theme or subject. As compared to shots the scenes have a superior semantics level. With the help of the similar content in the reminiscent component of semantics the documentation/segmentation of the scenes has been assembled by the successive shots. Images, manuscripts and acoustic way in the video have been that kind of evidence on which the federation have its foundation. The scene segmentation methods have been categorized according to the representation of the shots in three groups which have been defined subsequently: visual and auditory information foundation, key frame based, approach based on the background.

1) Ocular and aural information based: with the help of the subsequent approach a shot boundary have been designated in it the innards of pictorial and audio alteration occurs on the similar occasion in the procedure of the scene of the border. visual plus acoustic divisions have been recognized by the Sundaram and Chang [47]. Algorithm named time constrained the adjacent neighbour is used for the association of the willpower the scenes set in collaboration. The restriction is the visual and acoustic integration based approach have been that the founding of the relative amongst acoustic division and graphic reports are quite problematic.

2) Key frame based: Every shot of the video is identified as the subsequent in the form of the significant frames via them structures are occupied. The structures have been gathered temporally in the close shot act. Hanjalic *et al.* compute the resemblances between reports with the help of distinguishing the main frames [48]. Similar reports are being connected by linking the overlying relations in which the scenes have been divided. The gesticulation of paths are being mined as well as investigated after that they have been prearranged in the procedure appearance of capacities at the slices of sequential. For the capable characterization of the contents shot an assortment of gesticulation centred main frame are being used. In the shots of the adjacent main frame the section fluctuates, this is noticed by gauging its resemblance. The significant tactic of restrictions have been that main frames have not been capable of professionally display complete scopes of insides of the reports as in shots the scenes are frequently linked with the magnitudes of the insides in the act somewhat to the shots on the bases of frame main resemblances.

3) Background Based: The main theme about this approach is background similarity of same shots. Chen *et al.* a mixture of methodology are being used by them for the rebuilding of each solitary frame of audio-visual [49]. The background images of a shot got the approximation of the colour and circulation of texture for the conclusion of the resemblance of the report and the rules of film making are being used for the course of the federation shot procedure. It is the hypothesis of the contextual dishonourable method borders that the upbringings remain alike in the reports of the similar acts nevertheless occasionally credentials have been dissimilar in solitary act of the scenes shot.

Existing segmentation of the scenes have the dividable methodologies rendering to the dispensation technique, it is classified in four: based on splitting, based on merging, boundary shot and statistics based on a model.

a) Splitting based approach: the tradition of maximum to minimum flair theses tactic splits up the complete audio-visual in the formation of discrete scenes of intelligible. Rasheed and Shah [50] specimen elucidates it; for the video they are constructed in the similarity of the shot of the graph plus it has been used in he the regularized incisions for the division of the diagram. In each video of the scene there has been the depiction of the deputized diagrams. For the descriptive films the scheme is obtainable for the purpose of the group associated in the scene of the shots by Tavana pong and Zhou [51].

b) Merging-based approach: for the formation of the scene in the nethermost to the uppermost method it progressively merges the shots which are matching. Binary segmentation pass of the algorithm scene is being endorsed by Shah and Rasheed [52]. The consistency usage of the contextual shot terminated segmentation of the scenes has been initially occupied in the pass. The usage of the analysis motion at the second level the scenes which are segmented have been recognized and in the end they are emerged. For scene segmentation the best first model of the merging procedure was projected by Zhao *et al.* [53]. By HMM left to right algorithm in the consecutive shots it acquires each shot as a covered position and on the loops of the boundaries.

c) Classification based approach shot boundary: For the organization of limitations shot in the scene and non-scene limitations, the features of extracted shot boundaries the features of shot boundaries in this approach. A type independent method is offered by Goela *et al.* for the detection of limitations scene into videos broadcasting [54]. Act modification and non-scene change on these two sequences the scene portion is based on in their method. The shot boundaries have issues from SVM for the cataloguing. For the SVM, optimistic and undesirable preparation samples are being accomplished by gauged labelled

video scene parameters via innumerable genres broadcasts. The contrasts between the dissimilar reports have been exploited to pucker the matching shots in scenes.

d) Model based statistical approach: for the scenes division the method figures statistical models of shots. Stochastic Monte Carlo specimen is being used by Zhai and Shah for the stimulation of the scene [55]. With the previous step estimation the boundaries scenes are rationalized by dispersing, excruciating, and integration boundaries of the scene. The Gaussian Mixture Model (GMM) is used by Tan and Lu for the shots video bunches in acts rendering to separate shots structures [56].With density of the Gaussian, respective scene has been modelled. The combined diminished energy outline is being definite by Gu *et al.* In the restriction universal contented among solitary shots and the controlled local sequential neighbouring among shots is accessible. The ideal limitations divisions have been obvious by the elective process boundary [57].

Main worldwide in the fact approaches such as integration, arithmetical based model, and excruciating based which is impulsive and straightforward. In these methods the designated important frames set has one of the tactics that exemplifies these shots, although the representation of its active content shots. In its penalties both shots have been measured as a corresponding in that case if their main frames have been in the environment which is comparable, or if visually they have a similarity. The benefit of the native evidence concerning borders shot which are occupied by the boundary shot organization the constructed method. And it is approved that the algorithms which are in low intricacies they have accessible quality. The correctness of segmented scene decreases for the reason that the expectable have absence of universal knowledge of shots. The specific video characteristics domains are recognized, these domains are pictures/movies, newscast transmissions, as well as the TV is being used by the maximum existing methods for section subdivision,

for example, the rules of production usage with it TV indications arrangement have been completed [58], [59], [60]. The accurateness of the scene segmentation is being enhanced though the priori prototypical structure has become significant for each submission.

2.3 Extraction of Feature

The feature extraction accordingly to the structural analysis of the video have been the core of the video retrieval and indexing. The visual features were the researcher's focus on the audio-visual indexing as well as the retrieval. It consists of features of the main frames, gestures, as well as the substances. There have been no structures of transcript and aural assemblies.

2.3.1 Key Frame's Static Features

On the level of the appearances the audio-visual has been reproduced by the main frames of audio-visual. On the way to acquire key frames in audio-visual repossession, conventional techniques for carbon copy recovery have been implemented. The features of the static key frames chief cataloguing structures which have productive video retrieval and indexing have been created on colour, based on outline, as well as the base of the quality.

2.3.1.1 Features Based on Colour

The mixture of Gaussian models, the colour histograms, instant colour, coral grams colour etc. have been in the structures based in colour. The feature of colour based extraction have been reliant on the places of colour for instance, the (Hue,

Saturation & Value) HSV, (Red, Green, Blue) RGB, YCBCR, (Horizontal Vertical Colors) HVC and regularized r-g and YUV. Colour space choice is dependent on the applications. It is possible to extract the feature of colour from whole image or from set of images that are portioned into a complete appearance. Colour based features are the most effective features for video retrieval and indexing. Precisely, the colour histograms are suitable for informal descriptors. Colour histogram for the repossession of the audio-visual and notion detection are calculated by Amir *et al.* [61]. Initially, the images were sliced into 5 by 5 chunks to hold in the development of the colour by Yan and Hauptmann [62]. Afterward, for retrieval of the video, for each, histogram colour, moments of colour, and block have been taken out. Adcock *et al.* [63] used correlograms colours for the purpose of the filmed search engine grafting. The human optical view is imitation feature of the colour base because they are modest in abstraction and its extraction computational complexity has been squat; these are also its potentials. Texture is not defined in openly and shape etc., in the colour based features; this one is its disadvantage. Thus, the applications where shape and texture are essential, the colour based features are not effective.

2.3.1.2 Features Based on Texture

They are relevant to surface owned to object visual built-in features; they are autonomous in intensity or colour, as well as in images, they reflect the homogeneous phenomena. About the organization of object surfaces, they got very important information, along with that, its association with the environment around it. Simultaneous autoregressive models, co-occurrence matrices, wavelet transformation-based texture features, tamura features, orientation features are included in frequent use of texture features. Amir *et al.* used tasks for video recovery in TRECVid 2003 including co-occurrence texture as well as Tamura features using contrast and coarseness [61]. Gabor wavelet filters have been used by Hauptmann *et al.* [64] for the video

search engine by capturing the information of texture. Twelve oriented energy filters are designed by them. The filtered outputs of the mean and variance have been concatenated in a feature texture vector. The image was divided by Hauptmann *et al.* [65] in the blocks of 5 by 5 and the texture features were computed in each block, by using the Gabor wavelet filters. Texture based features benefits are that it is applied effectively on those applications which have information of texture as a salient feature in videos. Although in non texture video images, such features are not available.

2.3.1.3 Features Based on Shapes

From object contours or regions, the shape based features which define the shapes of an object in the image, can be extracted. Generally, the technique of detecting edges in the images is used and after that by using histogram the edges are distributed. The Edge Histogram Descriptor (EHD) has been used for capturing the spatial edges distribution, for TRECVid-2005 the video search task, by Hauptmann *et al.* [64]. The EHD, according to its quantized directions, has been computed with counting the pixels numbers which contribute to the edge. Images were divided in 4 x 4 blocks by Cooke *et al.* and Foley *et al.* [66]; to capture features of local shape, after that, for every block, edge histogram has been extracted. The applications in which the information shape is the main feature in videos, for them the shape based features are effective. Although, as compare to colour based features the extraction of above mentioned feature is very difficult.

2.3.2 Features of Object

Texture, size and the dominant colour etc., related to the objects of the image regions, are included in features of an object. To retrieve the videos which have the

similar objects, such features can be used [67]. In lots of retrieval video systems, the faces are the constructive objects. For instance, a person retrieval system has been constructed, that has the ability to get the shots of ranked list, which have a specific person and a query face in a shot has been given, by Sivic *et al.* [68]. A method has been proposed by Le *et al.* [69] with the integration of temporal information by converting into facial intensity information, in broadcast news videos for the retrieval of faces. To assist and comprehend contents of video, the texts in the video have been extracted as single object type. By increasing the semantics of a query as well as by using the Glimpse method of matching for the approximate performance of matching rather than matching it exactly, the text based video indexing as well as retrieval has been implemented by Li and Doermann [70]. Objects identification in videos has been very time taking and difficult as well, this is the drawback of object based feature. Rather than identifying various objects in various scenes, the current algorithms focal point is to identify the specific kinds of objects for instance faces.

2.3.3 Features of Motion

The distinguishing factor from the still images is the motion; it is the most important feature of the dynamic videos. By temporary variations, the visual content is represented by the motion information. As comparing to static key features and object features, the motion features were near to the concepts of semantics. In the motion of the video, the motion background is added, which is formed by camera motion as well as the foreground motion this is formed by the objects which are moving. Hence, video retrieval could be divided in two categories for motion feature based they are as following: object based as well as camera based. For video indexing, the camera based features and camera motions like: the in and out zooming, left and right panning, and up or down tilting are used. The limitation of video retrieval is by using the camera based features, that the key objects motions are not describable.

In modern work, a lot of interest had been grabbed by motion features of object based. Statics based, trajectory based, and spatial relationship based objects are the further categories of object based motion features.

2.3.3.1 Features Based on Statistics

To model the distribution of local and global video motions, the motion's statistical features of frames points were extracted in the video. Such as, the casual Gibbs models have been used for the representation of the distribution of the spatio-temporal for the local measurements related motions, which is computed after balancing, in the original sequence, the leading motions image, by Fablet *et al.* [71]. After that, for video indexing and retrieval, the general framework statistics has been adopted. The motion vector field has been transformed by Ma and Zhang into the directional amount of a slice in accordance to the motion's energy [72]. The set of moments has been yielded by these slices, for the formation of multidimensional vector known as texture of the motion. For the motion based retrieval of the shot, the texture of the motion is used. Statics based features extraction is low in complexity of computation; this is its advantage. Although its drawback is that the object actions cannot be represented perfectly as well as the link among objects can't be illustrated.

2.3.3.2 Features Based on Trajectory

In videos, with modelling the motion trajectories of objects, the trajectory features based were extracted [73]. An on-line video retrieval system has been offered by Chang *et al.* [74], it supported the spatio-temporal queries and object based automatic indexing. Algorithms for the video automated segmentation of the object and tracking are the part of the system. A trajectory based motion indexing compact as well as the proficient retrieval mechanism for the sequences of the video has

been presented by Bashir *et al.* [75]. By sub trajectories of temporal orderings the trajectories are represented. With the motion model, principal analysis coefficients components of the sub trajectories were represented. For the segmentation of the trajectory and to create, on velocity features, an index based wavelet decomposition was used by Chen and Chang [76]. Motion model was introduced by Jung *et al.* [77] for curve fitting of polynomial. For the individual access, the model of motion is used as a key indexing. To generate information in the form of the trajectory, for recurrent motion information and build from motion vectors a motion flow which was embedded in MPEG bit streams by Sue *et al.* [78]. A set of similar trajectories by given trajectories have been retrieved by the system. The trajectories were divided into many small segments as well as every segment was defined with a semantic symbol by Hsieh *et al.* [79]. For video retrievals; trajectories are matched by a distance measurement along with the exploitation of edit distance as well as visual distance. Object actions are describable; it is the advantage of the trajectory based feature. Its disadvantage is that, on correct object segmentation, as well as the trajectories automatic recording and tracking, its extraction is dependant and these are very difficult tasks.

2.3.3.3 Relationship Based Objects

Among the objects such features explains the spatial relationship. For the video retrieval application, Bimbo *et al.* [80] described the link among the objects by using the symbolic representation scheme. By the expression of every object, the arrangements of several moving objects and the specification of the spatiotemporal relationships were query by Yajima *et al.* [81]. The relationship among the objects of several types in the temporal domain can be represented spontaneously this is the relationship-based features objects advantage. Object and position labelling is difficult this is the drawback of these features.

2.4 Video Annotation

Various prerequisite concepts of semantics like car, person, people walking, and sky are the video segments and these are the shots for the allotment of the video annotation [82], [83]. The classification of the video is different in the ontology of category or concept than video annotation though few ideas can be applicable on both. In video annotation video shots or segments are applied where as video classification is applicable to complete videos, these are the only two differences among video annotation and classification, although they are quite similar. These are the following analogous methods of video annotation and video classification: foremost, the low level features have been extracted than different classifiers have been modified to chart the concept or category of the features labels. Video can be interpreted in different concepts with correspondence of its facts. Consequently, the reality of the video annotation with numerous concepts, they can be defined in the annotations of concept based, context based and integrated based [84].

2.4.1 Annotation Based on Isolated Concept

In a visual lexicon, this procedure of annotation has been used for every concept as a statistical detector trainee. For the discovery of multiple concepts of semantic, the classifiers of isolated binary have been utilized separately and autonomously although co-relation among the perceptions has not been measured. The distribution of multiple Bernoulli has been used by Feng *et al.* [85] for the image and video annotation sculpturing. The focal point of this model of multiple Bernoulli is clearly on the word annotation presence and absence, although it is done with the postulation that every word is independent than other words in an annotation. For every concept, the accuracy of different classifiers like GMM, HMM, kNN, and Adaboost have been inspected by Naphade and Smith [86]. For the performance of video annotation

semantic Song *et al.* [87] bring in the active learning together along with the semi supervised learning. A number of two class classifiers have been used in this method to take out of it with multiple classes of this classification. Depending on the construction of effective midlevel representations for the performance of classification of video semantic shots for sports video Duan *et al.* [88] utilized supervised learning algorithms. To load up the concept detectors in a single discriminative classifiers as well as to hold the errors of classifications which happen when in the feature space classes extend, a strategy of a cross training has been given by Shen *et al.* [89]. The link among the different concepts has not been sculptured and this is the restriction in the isolated annotation of concept based.

2.4.2 Annotation Based on Context

By using different contexts for different concepts the concept detection performance can be improved [90]. By using the context based concept fusion approach the context based annotation purifies the results of detection of the binary classifiers individually or concepts of deduced higher level concepts. With the previous reference, an ontology learning based procedure has been used by Wu *et al.* [91] to find the video concepts. To make the accuracy of the detection of the individual binary classifiers up to the mark, ontology hierarchy has been used. Model vectors have been made by Smith and Naphade [92] which is dependent on the scores detection of those classifiers which are individual; it is used to mine the not known and co-relations which are not direct among the prcised ideas, after that an SVM has been trained for the purpose of purgation of the detected results of an entity. For the annotation of the video Jiang *et al.*[93] introduced the active learning methodology. Users annotate some concepts in this method for few numbers videos and then these annotations are incorporated to deduce and develop other concepts of detections manually. With the help of unverified method of clustering an algorithm has been

made, which utilises enhanced pictorial ontology for the execution of the automatic video annotation of the soccer. The amounts and actions have been linked by default to the upper class ideas with the help of looking at the visual concepts proximity which has been hierarchically the part of the semantics of the higher level. For the training of the hierarchical classifiers of the video which have the sturdy relation among the video concepts, an enhancing hierarchical method, which has ontology concepts and multi-task learning, has been proposed by Fan *et al.* [94]. The individual detection has never been steady due to the detection of the miscalculation of the individual classifiers which can spread the fusion steps and because of that the division of the training samples occurs, which are for individual detection and conceptual fusion, correspondingly, due to the usual complexity of the correlations among the concepts there have not been enough conceptual fusion samples, and this is the context based annotation drawback.

2.4.3 Annotation Based on Integration

The following model covers the concepts of individual and its correlation at the same time. Concurrently, learning and optimization have been done. Along with that, all the samples have also been used for the individual concepts modelling and its correlation at the same time. A new feature vector is constructed, which grabs the concept's characteristics and concept's correlation, with the help a correlative algorithm of multi label proposed by Qi *et al.* [84]. The high computational complexity has been the drawback of the integration based annotation. The accurate amount of labelled training samples is required for the effective learning as well as robust detectors and with the help of feature dimensions the required numbers enhance exponentially. For the incorporation of the non-labelled data, few methodologies are proposed to convert it in the supervised process of learning to minimize the burden

of labelling. The classification of these approaches can be based on semi-supervised and of active learning:

1) Semi-supervised learning: The samples which are not labelled for the augmentation of the information for the available labelled models are used by this approach. To detect co-training based video concepts and for the investigation of the different strategies of labelling in co-training, which includes non-labelled data and few numbers of labelled videos, the semi- supervised cross feature for learning has been presented by Yan and Naphade [95]. By using small numbers of samples to learn concepts, Yuan *et al.* [96] proposed the algorithm of a feature based assortment containing multiple ranking. They comprise of three main components: pre-filtering, multiple ranking and feature pool construction. A video annotation algorithm has been proposed, which is based on learning which is semi-supervised with the help of kernel density approximation [97]. To handle the insufficiency of the training data in video annotation the semi-supervised learning algorithms based upon the optimized multi-graph has been proposed by Wang *et al.* [98]. The partially supervised learning system for the adaptive learning of different forms of objects and events for the specific video has been given by Ewerth and Freiskeben [99]. For the feature selection and total classification Adaboost and SVM has been included.

2) Active learning: it is a very useful and quick way for the handling of the lower sample brands. For the annotation of the video which comprises of the multi complementary predictors and adaptation of the increasing model, the algorithm of active learning has been proposed by Song *et al.* [100]. Along with that, it Song *et al.* [101] gave a framework of the video annotation proposal which was specifically designed for the personal databases of videos and it comprises of semi supervised and an active learning assembled process.

2.5 Video Indexing

The methods of video indexing are pigeon-holed on the bases of two main categories of contents, a) semantic (high level annotation) and: b) Perceptual (Low-level) [13][102] [103] [104]. Key advantages of video indexing based on perceptual features are:

- By using feature extraction techniques such as image and sound analysis, it can be fully automated.
- Users can search the similarities by using particular feature qualities such as the volume of the sound track and the shape and the colour of the object on the background.

However, users search videos which are semantic instead of low-level features, in contrast to; indexing based on features, this aims neglecting the semantic contents. There are some factors other than a perceptual level that makes feature-indexing more annoying and counter factual. Such as, users cannot always explain qualities for particular objects that they want to see do each inquiry. The support of more natural, powerful flexible ways of inquiring are some of the benefits of the high-level semantic based indexing. For instantce, semantic videos can be browsed by the users like topical classification. Moreover certain videos can be searched in consonance with the keywords. It is not easy to map the perceptual features to high-level due to the presence of semantic gaps, hence manual intervention is normally used to attain such indexings. Manual semantic annotation should be kept at minimum level due to its lengthy, biased and incomplete nature [105] [106] [107].

Three major indexing techniques are discussed in the subsections below:

1. Feature-based video indexing techniques, including shot-based, object-based, and event-based indexing
2. MPEG-7 based indexing
3. Annotation-based video indexing

2.5.1 Feature Based Indexing

Techniques of feature-based indexing can be categorized on the bases of extracted features and segments.

2.5.1.1 Segment Based Indexing Techniques

Subdividing a document into paragraphs, words, phrases, numerals and letters for distinctive sections during text based indexing where indices can be built on these constituents [108]. Video can also be separated into a hierarchy similar to the storyboard in filmmaking, using the similar approach [108] [109] [110]. For example, a hierarchical video browser is based on different level abstraction to help user in logically finding certain video farms or segments. This sort of arrangement is usually called the story board as it contains accumulations of frames, that belongs to main contents and events of the video. Less storage is required then whole video in storing key frames, so it reduces the bandwidth and delay requirements to render the video contents over a network for reviewing and browsing [102].

Figure 2.4 shows storyboard hierarchical video browsing indexing which can be explained below. A video can include shots from holidays, birthday party and a wedding ceremony. Every single of the stories is based on a set of scenes, for example a wedding story is based on the church blessing and the wedding party background. Every background is then further divided into shots, for example, different shots of

the special guests of the party and the exhilaration of the brides can be made up of the wedding party scene, whereas shot is made up of a sequence of different frames. Therefore, in corresponding indexing framework, story level gives a conceptual level of segment part, scene can be categorised as collection of distinct shots, shots can be classified as combination of different frames whereas frame is considered to be simple picture, have defined that a frame is a single image/picture, a consecutive order of frames is called shot, scene is a sequence of shots that correspond to a semantic content, and a story is a sequence of scenes that reveals a single semantic story. A deviation of this hierarchy is proposed in [111] where a video document normally has one or more purposes such as entertainment and information. A video document contains a genre of sports which contains many sub-genres like cricket, basketball and table tennis to achieve its purpose. Named events such as Six runs and out are the logical units of sports video are playing and break.

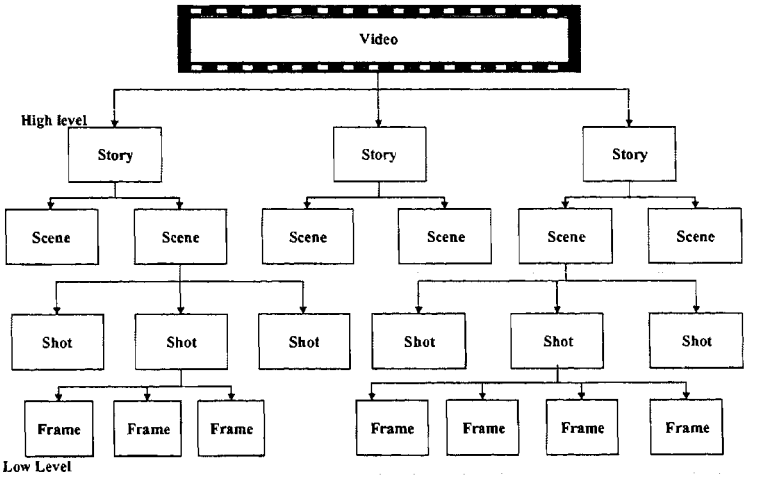


FIGURE 2.4: Segment based indexing

Key Frame Video segments can be represented in various ways; the most common type of representation is extracting the key frames. Key frames can be classified as most 'summarized' frame, which depicts the main idea of the shot. Key frames work just like key words in the textual domain. With the key frames, an image can be

co-ordinated during query or search for measuring similarity distance. Therefore, selection of appropriate key frames is very important. There are several methods for extraction of key frames automatically. The most common concerns in nominating the key frames are:

- The mostappropriate key frame selection
- Total number of key frames in a video

For the predecessor issue, it is commonly hard to automatically nominate the frames with the maximum semantic usefulness. This issue can be resolved by minimizing the frame which is redundant. This can be accomplished by adopting different methods, for example, relevance ranking (RR) [112] or singular value decomposition [113]. Another way of doing this is by sewing a frame sequence into a motion or mosaic icon [102]. The second issue can be overcome by allocating keyframes based on the length of video i.e., the selecting keyframes every second or two for lengthy videos [102]

2.5.1.2 Object-based Video Indexing Techniques

The main aim behind object-based video indexing is to differentiate specific objects in a whole video sequence to capture the changes in the content. Especially, a video with complex collection of objects is called a video scene, which is the physical characteristic and site of each object in addition to their relationship.
As compared to the extraction of perceptual features such as colour, volume and texture, the extraction of objects is much difficult. Video processing is a lot easier as compared to imaging as it changes an object section all together in the video frames. Furthermore, extraction of objects can easily be done when the video is compressed using object-based coding standards called MPEG-4. It interprets the video at a

high level of video contents which is better than prior standards of video coding e.g. MPEG-2 and MPEG-1. MPEG-4 is a group of one or more specific objects of video knows as video objects (VO) which contain layers of video object VOL(s). VOL contains systematic video object planes VOP(s). VOP is a semantic object consisting information of motion and shape in a scene. Yet, MPEG-4 only stipulates how to represent the objects in the compressed bit streams without requiring the information about how objects are detected and segmented, allowing the competition in developing object tracking and segmentation [102].

Object tracking is performed by detecting moving edges. A technique was proposed by Kim and Hwang [114] to segment VO based on edge change findings. They used Canny edge detector to identify and associate edges in sequential frames. Their output was a moving edge map which could extract VOP. The post processing involved rejecting unwanted moving objects. Contour-based tracking algorithm was developed by Schoepflin *et al.*, they used a sequence of template distortions to track and model the VO. The algorithm can detect the variations in the shape and pose of any objects by relaxing the constraints on the template distortions and renewing them sequentially after each frame [115].

It is very difficult to extract the objects in a frame when objects are many, very fast and small causing objects to blur. Therefore, an object-based indexing is not very effective in sport videos such as American football and soccer etc., because the players are many and moves continuously. Yajima *et al.* [81] introduced indexing framework allowing users interrogating object movement that including player and ball and querying by sketching a moving direction directly on a video screen. Their framework depended on drawing the spatio-temporal relationship among moving objects. This indexing technique was used by them as certain aspects of sport video cannot be retrieved and annotated by keywords. They did not use 'real' soccer video in their experiment and made soccer videos themselves containing few players, no complex background, no crowds and infrequent and complicated movements.

Additionally, the video was captured by one camera; eliminating the need to arrange panning, tiling and zooming of broadcasted soccer videos.

2.5.1.3 Event Based Video Indexing Techniques

Events in video segments can be detected by tracking object's activities. It aims to automatically detect the interesting events from raw video track [116].Yet there is a clear definition for 'event' in video indexing. The event is termed as the relationship between objects in a time interval that come before or after the next event [117]. An event contains:

a) Activities e.g. person walking: temporally periodic but spatially limited

b) Isolated motion events i.e. smiling: no repeat either in time or in space

c) Temporal textures like running water: unspecified spatial and temporal type.

Events in football are 'kick off', 'kick off return', 'punt return' and 'one side return' [116]. Ekin *et al.* [106] detected 'pass event' 'cross', 'shooting', 'score' and 'header', in soccer videos. Teraguchi *et al.* [118] introduced single of multiple triggering event(s) for the indexes e.g. 'kicking the ball' event triggered 'corner kick' or 'goal kick' , while 'receiving ball' triggered 'pass' event. In field and track sports, Wu *et al.*[119] detected 'standing up', 'still', 'throwing', 'walking', 'jumping up or down', and 'running' events in sports, to explore whether the event is a 'long jump' or 'high jump'. Furthermore, Lie and Sezan [120] examined 'plays' events in baseball, Japanese sumo and American football videos, constructing the smallest time-segment with necessary information. Similarly in cricket main events could be player getting out. Bowler bowling a no ball or player hitting a six. Some of the domain specific or general events can be categorized into generic and domain specific events as shown in Figure 2.5

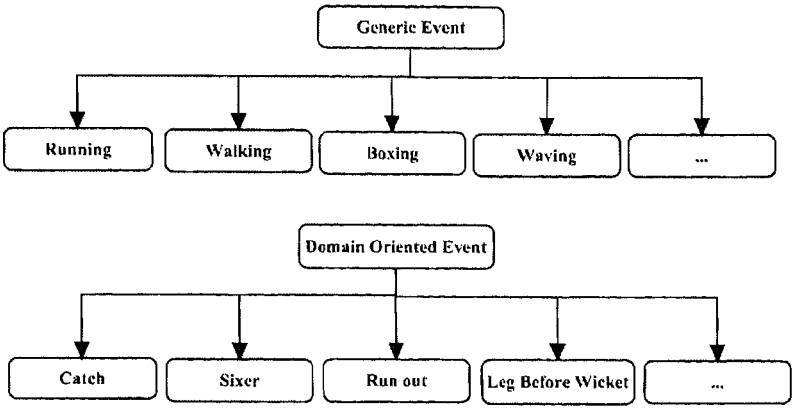


FIGURE 2.5: Generic and domain specific events

Event detection in sport videos is based on automatic analysis and manual work [118] of visual features by means of technologies e.g. camera or editing techniques [120], global motion detection [121], foreground-background extraction containing visible object recognition [119], and the recognition of CC (closed caption) streams [117]. These approaches for event detection are expensive and time consuming because they require numerous frames for analysis. Additionally, in sport videos, there is repetition in certain events that do not lead to a key event, e.g. passing, tackling and dribbling in soccer videos are common events, which happen many times during the match yet their detection would be a challenging and tedious process.
For sport videos, event-based indexing is more suitable as compared to object and segment-based indexing because:

- Players kicking ball and scoring goals are the specific events naturally decomposed by a sport match.
- Audience recall specific match on the bases of such events.
- In game videos, events can be summarised by certain audio-visual e.g., foul events can be denoted by referees whistle, text display and play being stopped.

- Events and their order of occurrence during can be automatically detected on certain knowledge of domain.

2.5.2 MPEG-7 Based Annotation Indexing

Moving Pictures Expert Group (MPEG) introduced MPEG-7 for the Multimedia Content Description as the ISO/IEC standard, which gives the description of many types of real-time as well as archived/non-real time audiovisual information. The large number of push (effective and fast browsing and search) and pull (filtering of data) applications can use it [122] [123].

An effort is required to optimize the usage of MPEG-7, because of the fact that recently proposed architectures of video indexing have used their own schemes of video indexing. Particularly, several alternative methods have already been considered by the MPEG-7 data scheme to illustrate video like graph, sequential and hierarchical models [124].The use of semantic graph can provide the organization of semantic annotations.

The aim of MPEG-7 is the only standardization of Multimedia Content Description Interface, in order to accommodate the competitive growth of the (automatic) feature extraction and filtering/ search applications. In order to achieve these objectives, MPEG-7 aims the standardization of the below mentioned components (which make the six elements of the MPEG-7 standard)[122] [125]

- ISO/IEC 15938-1: Systems, for the definition of architecture standard. As an instance, the scheme to prepare a MPEG-7 description to allow the synchronization between descriptions of contents and to achieve an efficient storage/ transport.

- ISO/IEC 15938-2: Description Definition Language (DDL), the standard language to define new or extending Descriptors (Ds) and Description Schemes

(DS). MPEG-7 DDL selected to implement XML Schema Language among various MPEG-7 specific extensions, in order to fulfill the requirements of MPEG-7.

- ISO/IEC 15938-3: Visual Ds, it consists of the fundamental structures as well as descriptors for the representation of the fundamental visual characteristics of multimedia contents. For instance, texture, colour, motion, location and shape.

- ISO/IEC 15938-4: Audio Ds, it holds audio descriptions of the multimedia contents. MPEG-7 audio consists of six technologies: framework for audio description, spoken language content, instrumental timbre, uniform silence segment, melodic-description tools and sound effect.

- ISO/IEC 15938-5: Multimedia Description Schemes (also called the MPEG-7 Multimedia DSs or MDS) intends the standardization of the descriptive tools set (Descriptors (Ds), Description Schemes (DS)) for the production of the multimedia data generic description (including text, visual, and audio).

- ISO/IEC 15938-6: Reference Software, also referred as an experimental model which is MPEG-7's imitation platform for Coding Schemes (CSs), Description Schemes (DSs), description Definition language (DDL) and Descriptors (Ds).

On the basis of MDS components, as presented in a Figure 2.6, this is eminent that MPEG-7 MDS illustrates multimedia data by its semantic aspects and structural aspects. This section is about the discussion focusing on the description of the MPEG-7 semantic data.

The emphasis of semantic DS is on the description of semantic entities like events, objects, semantic states, semantic concepts, semantic places and times in the narrative world. The narrative world is the framework where explanation makes the sense

that may cover the whole world described in multimedia data. Events and objects can be perceived and are considered as abstract entities which take place (or exist in space and time in a narrative world) within the multimedia data. On the other hand, semantic concepts are not the perceivable entities, or can be illustrated like the generalization of the semantic entities which are perceivable. Semantic states can be specified as the parameter attributes of relations and entities at a specific place and time, such as the age of an actor. Semantic places and times describe places and time in the narrative world respectively [126] [127].

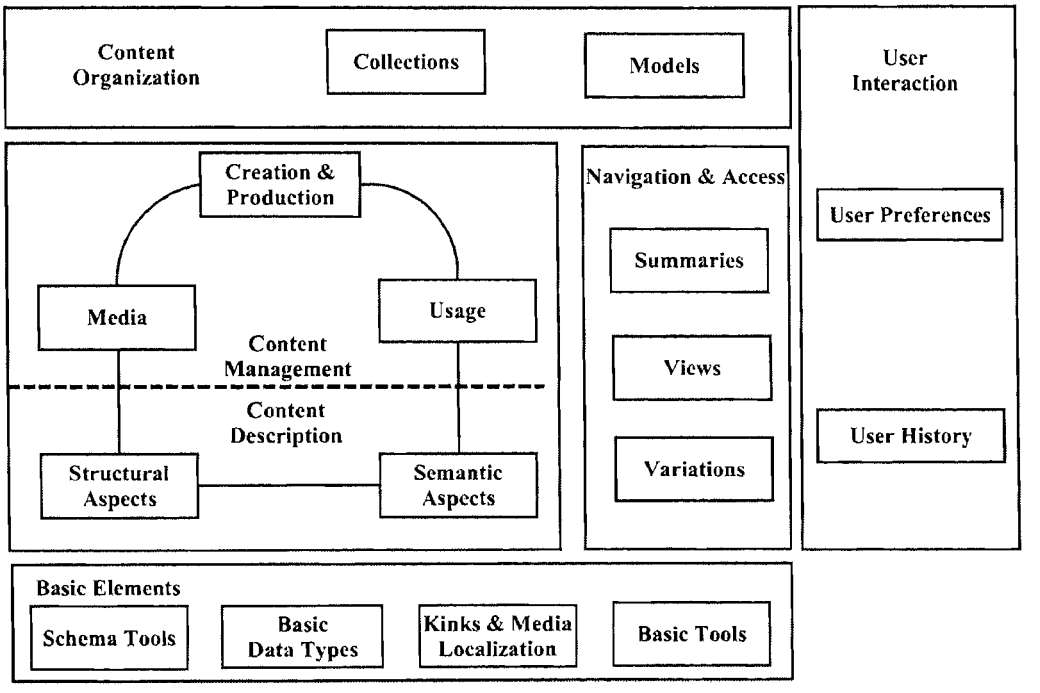


FIGURE 2.6: MDS components(MPEG-7)

Semantic DS may be used in forming the description of the abstractions that refer towards the procedure of description usage for the single instance to become generalized into the set of numerous instances of the multimedia content, or the pair of particular details. The abstractions are of two kinds, namely media abstraction and standard abstraction. The description separated from the particular instance of the multimedia content is referred as media abstraction. It gives a description of

multiple instances related to similar kind of multimedia content. For instance, how a description of the television news program can be utilized for the description of other television programs. The generalization of the media abstraction which gives a description of the general or common class related to semantic entities is referred as standard abstraction. The general method is the replacement of specific events, objects or other entities by using classes [126]. For instance, 'Obama running during his young age' may be substituted with "A young man running". Therefore, the standard description provides support for the instantiation of the description template.

2.5.3 Annotation Based Indexing

Management of text data like Information Retrieval techniques are well developed and supported successfully by traditional Database Management Systems (DBMSs). Annotation of the semantics related to video segments using free texts or keywords is, therefore, another alternative for the management of video. Thus, the standard query language like SQL can be used to manage user queries and hierarchical subject (or topic) classification can provide the basis for browsing [102], [128]. Thus, one key limitation of using such approach is that it can be extremely ineffective and tedious to annotate every video segment manually. Conversely, mapping the video features of low-level into semantic concepts of high-level is not the straightforward process. However, there also exist some major disadvantages that can be predictable using annotation based indexing. These may include:

- A selection of free text/keywords is subjective, which usually depends on the domain and application requirements.

- 'An image often worth a thousand words'. It means that a solo frame usually cannot be fully described by the words, thus it is assumed that words are not sufficient enough to describe the segment of video.

- It is usually the situation, when users are ignorant to explain using words what they are looking for, they will like to make a query based on the similar sound or image. In the same way users normally find the representation of visual key frames more interesting and helpful as compared to texts in a pure format.

In spite of these drawbacks, the exploration of this method is still required because of the fact that the annotations can be the nearest representation of video semantics contents. In addition, many keywords can be shared by the video applications like news and sports; therefore, to assist in making easier queries and making indexing more uniform, a 'glossary' can be created. The semantic graph MPEG-7 and subject classification are described in the next sub-sections as the instances of indexing based on annotation.

2.5.3.1 Subject (Topic) Classification

The implementation of topical or subject classification is mainly to organize huge collection of information like search engines and library. While browsing, users can select the topics which are available and the searching task can be executed on the basis of keywords for interested topics. There is a major limitation that the plurality of users has different view regarding the subject classification. Nevertheless, this constraint is not considerable because the users can make a search on the basis of keywords without surfing through hierarchies.

The video documents are organized in two levels. One is the subject classes for different videos and second one is the sub-classes which are used to extract video

shots. Though, further events and logical units or sub-classes can be formed by dividing each sub-class [102], [111].

The advantages offered by such indexing scheme are mentioned below:

- Several videos like movies and news have the topics which are well structured. However, the classification of sport videos is based on a sport type like individual and team, and events type.
- The contents of the document related to a single video like segments can be managed; moreover it is capable of organizing a huge video documents collection containing different topics.

The classification of video subject is presented by Lu [102] which splits the video into classes like food, art, travel, sport and animals. Snoek *et al.*, categorize the video document into information, communication and entertainment [111]. Talk shows can be classified into host and guest segments on the basis of logical units, while the sports like basketball and soccer contains break and play segments. Subject classification scheme demonstrated by Snoek *et. al.*, is easier to expand and more sophisticated than that of the scheme of Lu [102], because it is based on a thorough review of literature. Along with these approaches, the primary application of video can include sport, biomechanical investigation of sports, news, security, and feature films [128].The consideration of some issues is needed in order to classify every application. These topics may include:

1. Video content: typical or predictable contents
2. Video production: the way video is prepared with respect to filming, channel, composition and control over the script
3. Video usage: means of using a video

4. Video intent: an intention of making the video

On the basis of these classification schemes, Figure 2.7 represents the subject classification for a video which includes the genre and the sub-genres (which can further be expanded). For dividing them into events and logical units, the approach of domain-specific indexing can be applied.

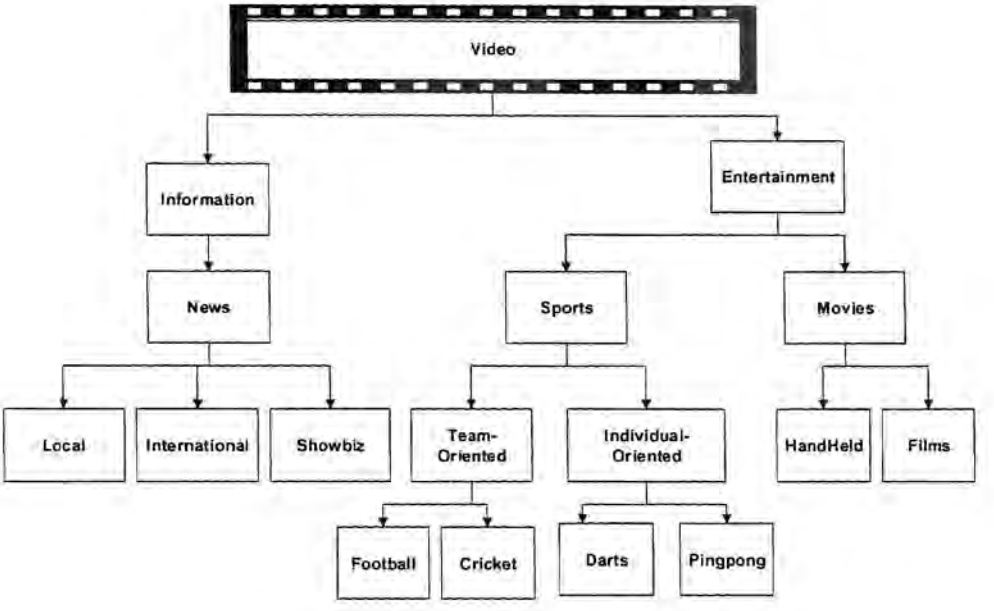


FIGURE 2.7: Video topic classfication

2.6 Query & Retrieval

The video retrieval which is content based starts its performance when the video indices have been attained. For the search of the user's video in accordance to the query sent by the user, the method of similarity measurement, which comprises on the indices is used. In reference to the feedback, the repossessed results have been optimized. Bellow are the similarity matching and the feedback relevance type queries have been reassessed.

2.6.1 Types of Queries

Those video queries which are non-semantic based, they are for instance: query by objects and query by subject. Those types of video queries which are semantic based are: query by natural language and by keywords.

a) **Query by Example:** From the sampled videos and images, this type extracts the low level features and with that by calculating the similarity of the given features the exact or similar videos has been found. For the query by example, the static key frames features suits best, as the stored key frames can be matched through the extracted key frames from the sampled videos and images.

b) **Query by Sketch:** By making sketches this query permits the user to get their desired video. The sketched features have been extracted and are matched with the stocked videos. A query by sketch method has been given by the Hu *et al.* [129] in it the path extracted from the videos are matched with the path sketched by the users.

c) **Query by Objects:** With the help of this query, a user can give the object's image. In the video database, all the events of the object are returned, when the system finds them [130].By comparing the example and sketch query, found outcomes of the are at the query object in the videos of query by objects.

d) **Query by Keywords:** It portrays the questions of users with the help of some key-words. This methodology is easy; it mostly gets the semantics and the main type of query. Video metadata, concept visuals and transcripts can be the part of keywords.

e) **Query by Natural Language:** The methodology of using natural language is the easiest way to make query. A semantic word comparison has been used by

Aytar *et al.*[131] for the repossession of the pertinent videos and then it divides them with reference to the query given with the natural use of language, mostly English. Parsing the natural language and attainment of exact semantics are the hard portions for the natural language.

f) **Combination-Based Query:** It is the amalgamation of different kinds of queries like text and video pattern based queries and is flexible for the multiple model findings. A frame work has been developed by Kennedy *et al.* [132] for the automatic finding of the grades of query with the help of query division in the training format with accordance to the different single model search method routine. The adaptive method has been given by Yan *et al.* [133] to mingle certain search gears for the application of the video retrieval query class dependency and in it the query class alliance weights of certain search gears are resolute automatically. Space query by person and without a person requires are differentiated by the multimedia system of retrieval. With the treatment of query classes as the latent variables Yan and Hauptmann [134] took the quries classification as well as combination weights determination into the structure of probabilistic.

2.6.2 Similarity Measure

In video retrieval, the video similarity measurement has the vital role. Combination matchings, text matching,ontology based matching and feature matching are methodologies for the measurement of the video similarities. On the type of query, it is depended to choose the method.

a) **Feature Matching:** The general space among the corresponding frames features is the accurate similarity of measurement in between the two videos. For the findings of the similar videos, query by example mostly incorporates the low

level features. In various granularities and resolutions video similarity can be taken [135]. For the measurement of the video similarity; ; the key frames for static features [136], object features [137] and motion features [129], are used in according to user needs. Sivic *et al.* [68], shifted the features from the sample shot which has to query face and then with the saved face features those extracted features were matched. After that the queried faces were retrieved, which were in the shots. In the set of videos, Lie and Hsiao [138] extracted the major objects trajectory features, after that for the videos retrieval it matched the extracted trajectory features with those trajectory features which were stored. In the space of the feature, the similarity of the video can easily be measured, this is the plus point of the feature matching. The negative point is that the similarity of semantic is not representable due to the spaces among the sets of vectors features as well as to people semantic categories are very similar.

b) **Text matching:** To find the video, name equivalence of every idea by its query is the easiest method and this method is also suitable for the query. With the help of vector space query text and text description were computed to find the similarity among them, before that both the concept description and the text query were regularized by Snoek *et al.* [139]. The most similar ideas were chosen at the end. Text matching approaches are instinctiveness and easy implementation are the pros of it. To get the appropriate results of the research complete matching concepts should be completely added in the query text, this is the drawback of this approach.

c) **Ontology based matching:** The matching of keyword semantic relation and semantic concepts resemblance ontology are done by this approach. From the knowledge sources like, concepts ontology and keywords, query descriptions have been developed. The words in the query text are disambiguated syntactically, after that noun chunks and nouns are translated by finding every noun from

word net; they are extracted from the concepts ontological concepts by Snoek *et al.* Because of the link of the concepts to the word net. Thus, to find the concept's relation to the actual text query is determined by the usage of ontology. On the facts based word similarity, semantic is a worthy visual co-occurrence approximate. For the similarity findings of the text annotated videos and quires of users, Aytar *et al.* [131] used semantic similarity word measures. By the help of text query defined by a user, the videos are retrieved depending upon its relevance. For the improvement of the retrieved results, the additional concepts are incorporated from the knowledge sources [140], this is the benefit of the ontology based matching approach. The concepts which are inappropriate are taken, which leads to the unwanted corrosion of the findings; this is the negative point of this method.

d) **Combination based matching:** The combination methodologies are learned by the training collections; such as a combination of independent query with learning models and mixture of class and query models which empowers the semantic concepts [62]. This approach is very much handy for quires based on combination, which has the multi modal searches flexibility. Combination based matching approach can automatically determine the concept weights as well as by far it can handle the hidden semantic concepts, these are its benefits. Query combination models are not easy to learn; this is its drawback.

2.7 Video Retrieval

Hampapur *et al.* [128] and Elmagarmid *et al.* [109] have proposed a classification scheme for video queries. They classified queries according to the following categories.

Query content - on the nature s of video content's which is necessary to accomplish the need for query, is categorized. Bases of certain categorisations are:

- Semantic (information) query: Its requirement is to recognise the semantic contents of high-level related to the data of video. The annotation information of semantic video can be used for its partial satisfaction, which reflected as a complex class of query on the bases of video database consisting of technology areas e.g. artificial intelligence, computer learning and machine learning etc. The examples include: Emotion equals 'Laughing' and searching scenes containing Actor equals 'X'.

- Audiovisual (AV) query: it requires video's AV contents and does not rely on the understanding of semantic data. AV queries are fulfilled after automatic or semi-autmated analysis e.g: shots search where camera is immobile with zoom-in lens.

- Meta (information) query: Its responsibilities are the extraction of the information of video data e.g. production date, producer and the length are the chores performed by certain queries. Usually it is contented with straight texts that are built on automatic searching, however there is a chance of getting inserted manually, online or offline for meta-annotations. For instance searching a video with its title or the name of director.

Extraction of similar kind of objects is gained by Query using matching from the data base. Therefore, in order to match the sample of video data and query on AV features , a fraction of processing is performed e.g. sound and image analysis to correspond the sample of video data and query.

- precise correspondance is needed between query andvideo data.e.g. a scene search for object "Actor X"

- Identical-match query (Often known as Query by Example): it is differentiated from Exact-match query because precise matching is prohibited in query by example. Due to its complex nature, this query is famous and commonly used as well e.g. exploring every shot where object appears like a given image.

Query behaviour or query functions: It is based upon certain functions executed by the queries.

- Browsing query: This type of query is used when the users are not familiar with the types and structures of information available in the database or they are not clear regarding what they are able to retrieve from that video database. Fuzzy queries formulation to browse within the database should be allowed by the system. It does not require some particular entities while browsing, hence a few data sets must be provided by the system for the representation of each video data present in that system, like key-frames or iconic descriptions along with some annotations of textual metadata. For instance: browsing the video contents with genre equals "Movie".

- Location-deterministic query: This type of query attempts to locate, in particular, some video information in the database; hence the exact idea about the video data is compulsory for the user. For instance: searching for all scenes for the location with actor 'A'. A pointer to the start of such scenes is returned as a result.

- Iterative query: This is considered to be an extended type of browsing query; since the graphical user interface (GUI) is provided in the system which offers users the ability to filter their queries incrementally until the results are generated up to the maximum satisfaction level.

- Statistical queries: This type of query needs statistical functions on the information from the single video or within whole database. For instance: to

calculate the percentage of videos present in the database on the basis of their genre, e.g. sport 15%, news 50%, etc.

- Tracking queries: This type of query is used to track the visual quantities present in video. For instance: tracking of a car throughout the scene. The outcome is the position of car in each frame in the scene.

Query temporal unit such queries categorize the video data granularity that is required to fulfil query.

- Unit (or video-stream based) query: : It controls the entire units of video e.g.: searching for a drama video with the writer 'A'.
- Sub-unit query: This type of query is distinguished from the Unit query in a way that it deals only with some part of video, like scenes, clips, and frames. For instance: searching the scenes in which the actor 'A' appears, searching for the shots where a person with 'such face type' appears.

2.8 Chapter Summary

In this chapter, we had discussed the in-depth survey in the context of video indexing and information retrieval technology. We had a comparison of existing work in the area of video indexing. Video is an important component containing certain features that make its retrieval and indexing complex. NIST and TREC are two main companies working on improving the video retrieval and analysing experimented videos to improve its quality by means of introducing different algorithms.Video segmentation is done on the bases of two main categories i.e. shot boundary detection and scene segmentation. Visual features are the main constituents for feature extraction, these are used for the key frame features, objects and motion features however acoustic

or text features were not part of it visual feature but could be processed separately. Semantic annotation is required to get the high level features such as car, people, sky etc. which normally classified in scenes detection whereas shots could be detected using low level features. Video annotation has different theories having domains such as concept based, context based and integrated based theories. Low level and high level features are two main levels for video indexing. Both have their own benefits in terms of video indexing. The video queries have certain kinds which are query by example, by sketch, by object, by key word, combination based etc. They have helped to match the text, features of similar videos depending on the type of query. A classification system for video queries comprises of framework where videos were categorised on the bases of semantic information, AV query, meta-data information, location, iterative, tracking and statistical queries.

Chapter 3

Framework for Dynamic Intra-video Indexing

The approaches discussed in literature tend to address the problem of feature based indexing in an atomistic approach focussing on individuals components of the process using available analytical and or stimulation techniques. These approaches can either be based on low-level features or high-level features. Main advantages of low-level features are: a) they are computational cheap and can be automated fully based on features such as texture or image analysis; b) User can make of characteristics of simple feature such as object to search using these features. Their performance may give plausible results individually, however, there is a need for integrating different component into a framework to evaluate the gradual systematic performance for video indexing and retrieval system which could be easy to implement and evaluate. The gap identified in the literature review leads to formulation of the video indexing framework to fulfil this distance between low level and high level features. The novelty of the framework will be its handling of the numerous components for feature extraction, video indexing and retrieval, which formerly has been treated separately. The constituents are enclosed in the most practicable sequence they seem in video

indexing and retrieval systems. The framework will also implement the retrieval phase particularly from the user's perspective.

To overcome the performance and evaluation issue as a whole we proposed a holistic framework entailing all the sub processes of video indexing from input till the final output. The schematic diagram of the proposed approach illustrate in figure 3.1

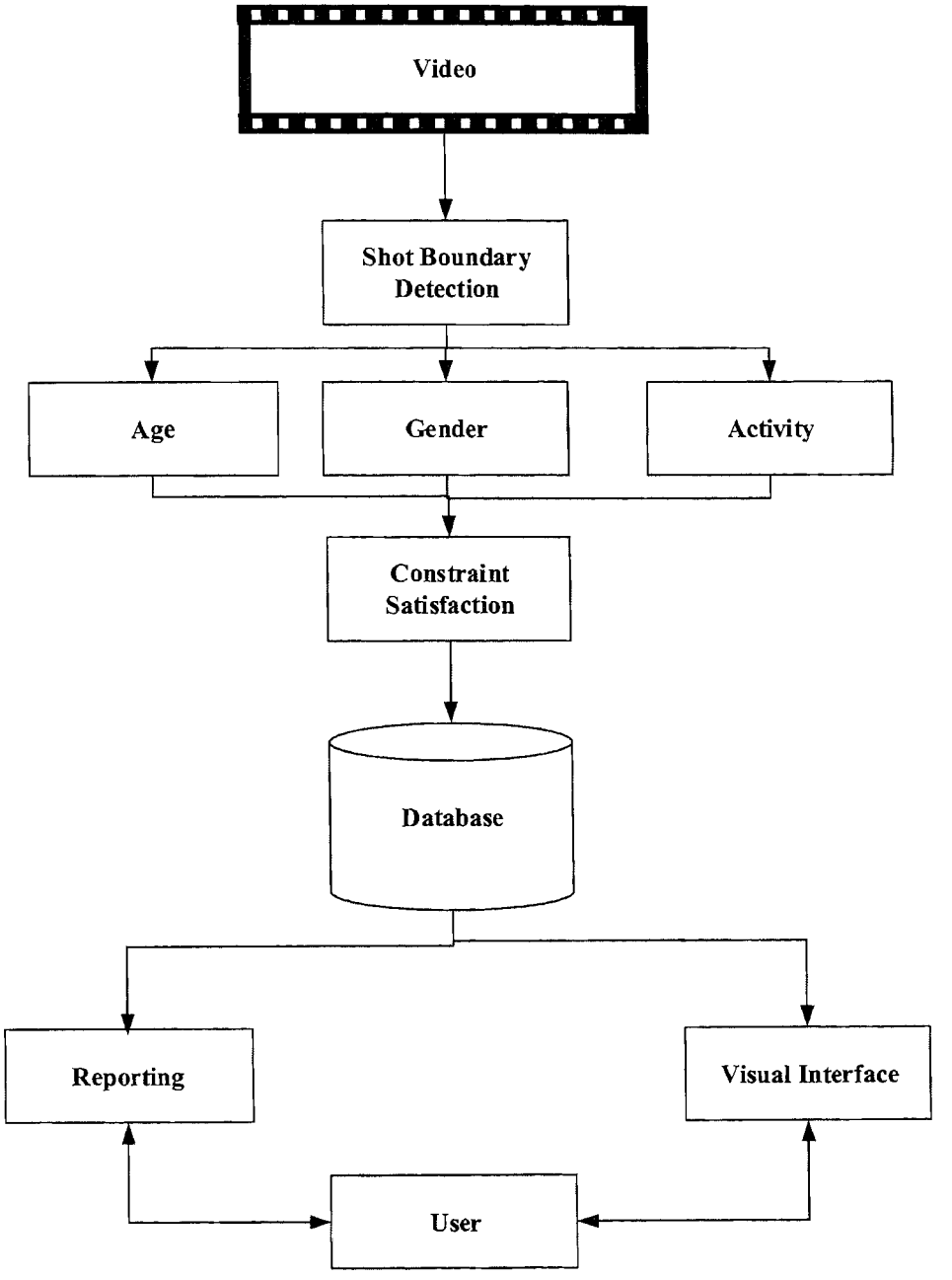


FIGURE 3.1: Proposed holistic framework

3.1 System Architecture

The architecture of the proposed system is depicted in figure 3.1 which covers the main components of a video indexing system. The proposed framework focuses on each and every step, keep in view of the whole cycle of video indexing system. The proposed framework can be categorised into three broader phases:

- Detection phase
- Indexing phase
- Retrieval phase

The detection phase starts with the detection of shot boundary detection to divide the video into chunks for indexing purposes. These chunks will be further divided into frames for extraction of features which will be used for detecting age, gender and activity in the shot. These features will be tested for constraint satisfaction and will be indexed based on the detection of features. These indices are then used for retrieval purposes and all the shots based on specific query requirements could be retrieved upon request from user. The videos indices will be used to support retrieval and browsing by use of a novel GUI.

3.2 Database Design & Querying

Database management system (DBMS) can only follow a database structure written in formal schema language supported by it. Database schema gives a blueprint of the structure of database, tables and columns; along with the expected data to be filled in it. The database schema also comprises of procedures known as integrity constraints to enforce compatibility among different fragments of the database which

are defined in the same language. A conceptual schema leads to explicit mapping which is transformed into the actual database. The schema also outlines tables, relationships, indexes, views, fields, queries, etc. which are normally stored in Data Dictionary (DD). Schema can either be defined as graphical representation or text based demonstration. In order to facilitate the user with accurate indices, a database has been designed to hold the location of indexes along with features extracted in feature extraction phase. Database comprises of four main tables. Anything one could possibly want to know about how these object was built is documented which generates the blueprint of the database. It is handy report and helps in sharing the construction progress with other researchers. A detailed text based schema of the database designed to be used this research is illustrated below:

1. tbl_Video (VID, VNAME, VLENGTH, VFORMAT, VFPS, VIDEOPATH)
2. tbl_Segment (VID, SNO, SST, SET)
3. tbl_Features (FID, FTYPE, FVALUE)
4. tbl_FinalResult (VID, SNO, FEATUREIDD)

Video table contains the basic information about the videos present in the dataset. VID is the primary key in this table to uniquely identify the video among other videos of the dataset. VNAME is the title of the video which will also be helpful in giving a naming convention to a numeric video id. VLENGTH contains the total length of the specific video in number of section. VFORMAT contains the specific video format, e.g, avi, mp4, mkv etc. VFPS contains the video frames per second which is different for different standards of the video. This will help in calculating the total number of frames in a specific video by multiplying video length with frames per second. VIDEOPATH contains the path of the video where it is being

stored. A detail documentation about the different fields along with their default values of video entity is given in table 3.1 using Microsoft Access ® 2013.

TABLE 3.1: Video table schema

Field Descriptor	VID	VNAME	VLENGTH	VFPS	VFORMAT	VPATH
AggregateType	-1	-1	-1	-1	-1	-1
AllowZeroLength:	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
AppendOnly	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Attributes	Fixed Size	Variable Size	Fixed Size	Fixed Size	Fixed Size	Fixed Size
CollatingOrder	General	General	General	General	General	General
ColumnHidden	False	False	False	False	False	False
ColumnOrder	Default	Default	Default	Default	Default	Default
ColumnWidth	Default	Default	Default	Default	Default	Default
CurrencyLCID	0	0	0	0	0	0
DataUpdatable	False	False	False	False	False	False
OrdinalPosition	1	2	3	4	5	5
Required	False	False	False	False	False	False
Type	Int	Text	Text	Int	Text	Text
Size	4	255	255	4	255	255
SourceTable:	video	video	video	video	video	video
TextAlign:	General	General	General	General	General	General

Segment table contains information of shot segments within a specific video. VID is the foreign key from table video. SNO is the segment number within a specific video. VID and SNO make a composite key to uniquely identify the segment number within a specific video. Any video can have different shot boundary segments so specific video will stored in the VID and the shot number will be stored in the SNO of that specific video. Also segment number and video id cannot contain a duplicate value, which means any shot number in a specific video can only come once. SST holds the segment start time of the shot in second whereas the SET contains segment end time of the specific shot. First shot default start time will be zero whereas last shot end time will be the total length of video in seconds. A detail documentation using Microsoft Access ® 2013 about the different fields along with their default values of segment entity is given in table 3.2.

TABLE 3.2: Segment table schema

Field Descriptor	**VID**	**SNO**	**SST**	**SET**
AggregateType	-1	-1	-1	-1
AllowZeroLength	FALSE	FALSE	FALSE	FALSE
AppendOnly	FALSE	FALSE	FALSE	FALSE
Attributes	Fixed Size	Variable Size	Fixed Size	Fixed Size
CollatingOrder	General	General	General	General
ColumnHidden	False	False	False	False
ColumnOrder	Default	Default	Default	Default
ColumnWidth	Default	Default	Default	Default
CurrencyLCID	0	0	0	0
DataUpdatable	False	False	False	False
OrdinalPosition	1	2	3	4
Required	False	False	False	False
Type	Int	Int	Int	Int
Size	4	5	6	6
SourceTable	Segment	Segment	Segment	Segment
TextAlign	General	General	General	General

Features table contains detailed set of features extracted during the feature extraction process. FID is the primary key to uniquely identify the feature. FTYPE contains the type of feature under consideration such as gender, age, activity. FVALUE contains the value of gender type. For example if FTYPE is gender, then its value can be male or female. Similarly age FTYPE can contain value baby, young and old. Thirdly activity can have walk, wave, run etc. FID given a specific mask of pre script for every FTYPE making it optimum for searching purposes. All these

together make composite key to uniquely identify a specific feature. A detail documentation using Microsoft Access ® 2013 about the different fields along with their default values of feature entity is given in table 3.3.

TABLE 3.3: Feature table schema

Field Descriptor	**FID**	**FVALUE**	**FVALUE**
AggregateType:	-1	-1	-1
AllowZeroLength:	FALSE	FALSE	FALSE
AppendOnly:	FALSE	FALSE	FALSE
Attributes:	Fixed Size	Variable Size	Fixed Size
CollatingOrder:	General	General	General
ColumnHidden:	False	False	False
ColumnOrder:	Default	Default	Default
ColumnWidth:	Default	Default	Default
CurrencyLCID:	0	0	0
DataUpdatable:	False	False	False
OrdinalPosition:	1	2	3
Required:	False	False	False
Type	Int	Text	Int
Size	4	255	4
SourceTable:	Features	Features	Features
TextAlign:	General	General	General

Final result contains indexes location within a video along with feature set. This is the main table to contain the primary information of the indexing system. Video id is the foreign key from video table which is reporting the information about the video under consideration. Segment number is the specific segment within the video

containing the specific feature. Feature IDD is the foreign of feature number which is present in the specific segment of the video under consideration. This feature id can be mapped to features table and the resulting feature value could be obtained. All of these constitute the composite key for this table which will be used to uniquely identify the specific video and its segment along with the features as required by the user. A detail documentation using Microsoft Access ® 2013 about the different fields along with their default values of final table entity is given in table 3.4.

TABLE 3.4: Final result table schema

Field Descriptor	**VID**	**SNO**	**FIDD**
AggregateType:	-1	-1	-1
AllowZeroLength:	FALSE	FALSE	FALSE
AppendOnly:	FALSE	FALSE	FALSE
Attributes:	Fixed Size	Variable Size	Fixed Size
CollatingOrder:	General	General	General
ColumnHidden:	False	False	False
ColumnOrder:	Default	Default	Default
ColumnWidth:	Default	Default	Default
CurrencyLCID:	0	0	0
DataUpdatable:	False	False	False
OrdinalPosition:	1	2	1
Required:	False	False	False
Type	Int	Int	Int
Size	4	5	4
SourceTable:	Video	Segment	Features
TextAlign:	General	General	General

Once tables has been populated the retrieval is performed using text based queries. An example for text based query in this research shown below.

```
SELECT DISTINCT Video.VID, Video.VName, Video.VideoPath, Video.VLength,
Segment.SST, Segment.SET
  FROM Features
    INNER JOIN Video
     INNER JOIN Segment
     INNER JOIN FinalResult
     ON Segment.SNo = FinalResult.SNo
     ON Video.VID = Segment.VID
     AND (Video.VID = FinalResult.VID))
    ON Features.FID = FinalResult.FeatureIDD
  WHERE FinalResult.FeatureIDD=SPECIFIEDID
```

Appendix 'databaseschema' contains the XML schema files for database designed.

3.3 Summary

In this chapter, we had presented the framework where the holistic architecture of the video indexing and retrieval system has been illustrated and defined components and structure wise. The presented framework bridged the gaps of the retrieval system which were discussed in the literature review related to indexing in an atomistic approach focussing on individual components of the process using available analytical and or stimulation techniques. In addition, the gap of formulation of the video indexing with respect to distance between low level and high level features has been overcome. To conclude, the novelty of the framework filed the gap in the area of

feature extraction, video indexing and retrieval, which formerly havebeen treated separately.

Chapter 4

Video Corpus Production

4.1 Introduction

The key step for the study is to create a dataset having all the features which will be evaluated by the system. Due to unavailability of any dataset for all the feature used in this research, a new dataset is created for this purpose. The main focus is the presence of human as human is the most important aspect in any video. The human focused dataset benefits in managing the domain of the research. The dataset will be used for annotation and evaluation purposes. Until now, dataset is being designed and manually annotated having a videos with different gender, age group and activities. Every segment of video in the dataset comprises of single and multiple camera shots mix together for identification and evaluation purposes. Human annotation will be used as a reference to the annotation generated by machine. Detail of dataset creation will be discussed in the section 4.2.

4.2 Video Corpus

In order to measure the true performance of the video indexing and retrieval system, the dataset plays an important role. It is crucial that dataset need to be created on standard set and should meet the baseline requirements of the particular domain. The face recognition dataset [141] used by many researchers comes with the faces in various settings and mainly used for face detection in an open environment. The dataset is mainly tagged for the same purpose, the data set provides a fair amount of data in a wild environment. The idiap/ETHZ dataset [142] is also taken as a baseline dataset the additional feature besides the face is the poses. The dataset is mainly used by the researchers for faces and pose detection; the key areas dataset could be used are face detection, pose detection, walking and the supporting text for learning. The Caltech dataset [143] produced by the computer vision group working at Caltech Institute is another prominent dataset specially for gender detection the data set comes with 267 male and 168 female faces. The ICG dataset [144] is another data set comes with the data specially in wild and annotated for face detection, pose action. Besides this the Yale face dataset [145], the CMU faces dataset [146] mostly used for frontal faces and Our Database of Faces (ORL)dataset [147] for face recognition , and identification of human expressions are the commonly used datasets. The table 4.1 highlights the salient features of the datasets which are commonly used for age, gender and activity detection.

The datasets discussed in table 4.1 are purpose built by various researchers to test and evaluate the performance of their algorithm for face detection and recognition along with other features. However, many of these dataset are domain specific and support certain features and provides little or no support for others. This raises the need of developing a dataset which could be useful for evaluating performance of video retrieval system specifically related to age, gender and activity detection.

TABLE 4.1: Existing dataset

Data set name	Year	Papers
CHALEARN Multi-modal Gesture dataset	2014	[148]
Shefeld Kinect Gesture (SKIG)	2013	[149]
Berkeley Multi-modal Human action Database (MHAD)	2013	[150]
MSR action	2012	[151]
CASIA Gait Identification Dataset	2011	[152]
MuHAVi and MAS human action	2010	[153]
Hollywood Movies	2007	[154]
Weizmann action	2005	[155]
KTH action	2004	[156]
POSTECH Labelled faces in the Wild	2013	[141]
ICG Annotated Facial Landmarks in the Wild (AFLW)	2012	[144]
Idiap/ETHZ Poses and faces	2009	[142]
Computational Vision at Caltech	2004	[143]
The Yale face	2001	[145]
CMU faces - Frontal faces	1998	[146]
Our Database of faces	1994	[147]

4.3 Video Corpus: VCAGA

The Video Corpus for Aaction Gender Age (VCAGA) dataset comprises of videos made indoor with static background having different volunteers. The dataset comprises of five types of actions performed by human (walking, running, boxing and hand waving, laying down) which have been performed several times by 41 subjects in a scenario which is indoors. At present the dataset comprises of 4395 sequences. All clips were filmed with a stationary camera having a homogeneous white background with 25 frames per second frame rate. All clips are stored using mp4 file format. The resolution of videos vary significantly as these are the standards which were either used by other researcher or could be used for research purposes. It varies from very high to very low keeping in mind the processing capabilities. Any researcher can use the dataset according to their requirements. Any classifications

can be performed keeping in mind the resolution restrictions of specific task. Classification could be uni-dimensional or multi-dimensional based on the requirements of research as well as the limitation of resolution. If the resolution under consideration fulfil the requirement of identifying different modalities then it can be used in conjunction where as if it fails then it can only be used to identify single modality. So the classification of either age, gender or activity can be performed one at a time or in conjunction based on algorithm developed for classification purposes.The clips were sampled at different video standard as shown in table 4.2.

TABLE 4.2: Video standards

Type	**Pixel Width**	**Pixel Height**
FHD	1920	1080
WSXGA+	1680	1050
HD+	1600	900
UXGA	1600	1200
WXGA+	1440	900
SXGA+	1400	1050
HD	1366	768
HD	1360	768
WXGA	1280	720
WXGA	1280	768
WXGA	1280	800
SXGA (UVGA)	1280	960
SXGA	1280	1024
XGA+	1152	864
WSVGA	1024	600
XGA	1024	768
SVGA	800	600
VGA	640	480
HVGA	480	320
QVGA	320	240
QCIF	176	144

4.4 Annotation Process

User based annotation is performed and all video sequences are annotated using the keywords based on age, gender and activity. Video clips are carefully annotated to make a data dictionary for the purpose of testing and validation. Each annotation is performed keeping in mind the high level feature which are presented in each clip. Along with the high level features, the key information about different aspects of the video such as shot boundary, total number of fames in the sequence, age of individual present in the shot along with its gender and the action it is performing is also annotated. Figure 4.1 shows categorization and annotation of human related information.

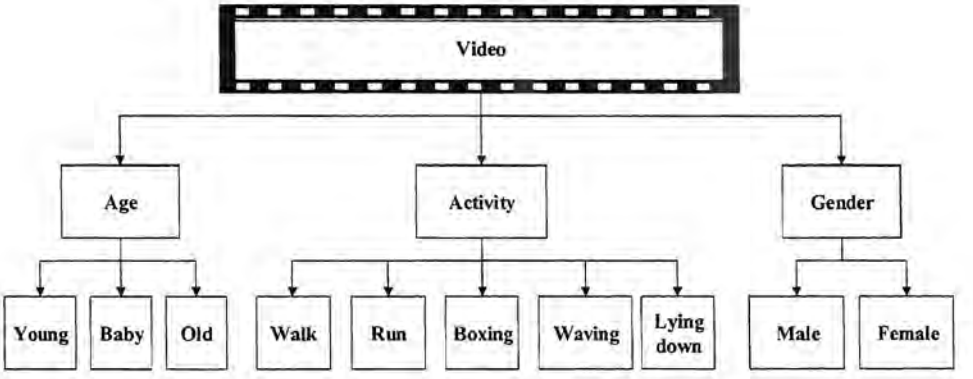


FIGURE 4.1: Categorization and annotation of human focussed videos

The annotations of the clips are prepared with open English vocabulary and use of special characters which are specific to computer are avoided to make it general for naive users. The name and addresses of volunteers were not used in the annotation process as all volunteers are anonymous. Also all the audio information is being discarded as the main focus is on the visual contents of the video.

Figure 4.2 exhibit a segment of the video, which is down sampled at 1 frame per second for illustration purposes, along with the annotation of a male young man waving.

FIGURE 4.2: Male, young, waving

Figure 4.3 exhibit a male young man who is running in the video along with the annotation.

FIGURE 4.3: Male, young, running

Figure 4.4 exhibit a male boxing along with the annotation.

FIGURE 4.4: Male, young, Boxing

Figure 4.5 exhibit a segment of the video, which is down sampled at 1 frame per second for illustration purposes, along with the annotation of a female young man boxing.

FIGURE 4.5: Female, young, boxing

Figure 4.6 exhibit a young man walking whose gender is male man in the video along with the annotation.

FIGURE 4.6: Male, young, Walking

Figure 4.7 exhibit a segment of the video, which is down sampled at 1 frame per second for illustration purposes, along with the annotation of a female young man walking.

FIGURE 4.7: Female, young, walking

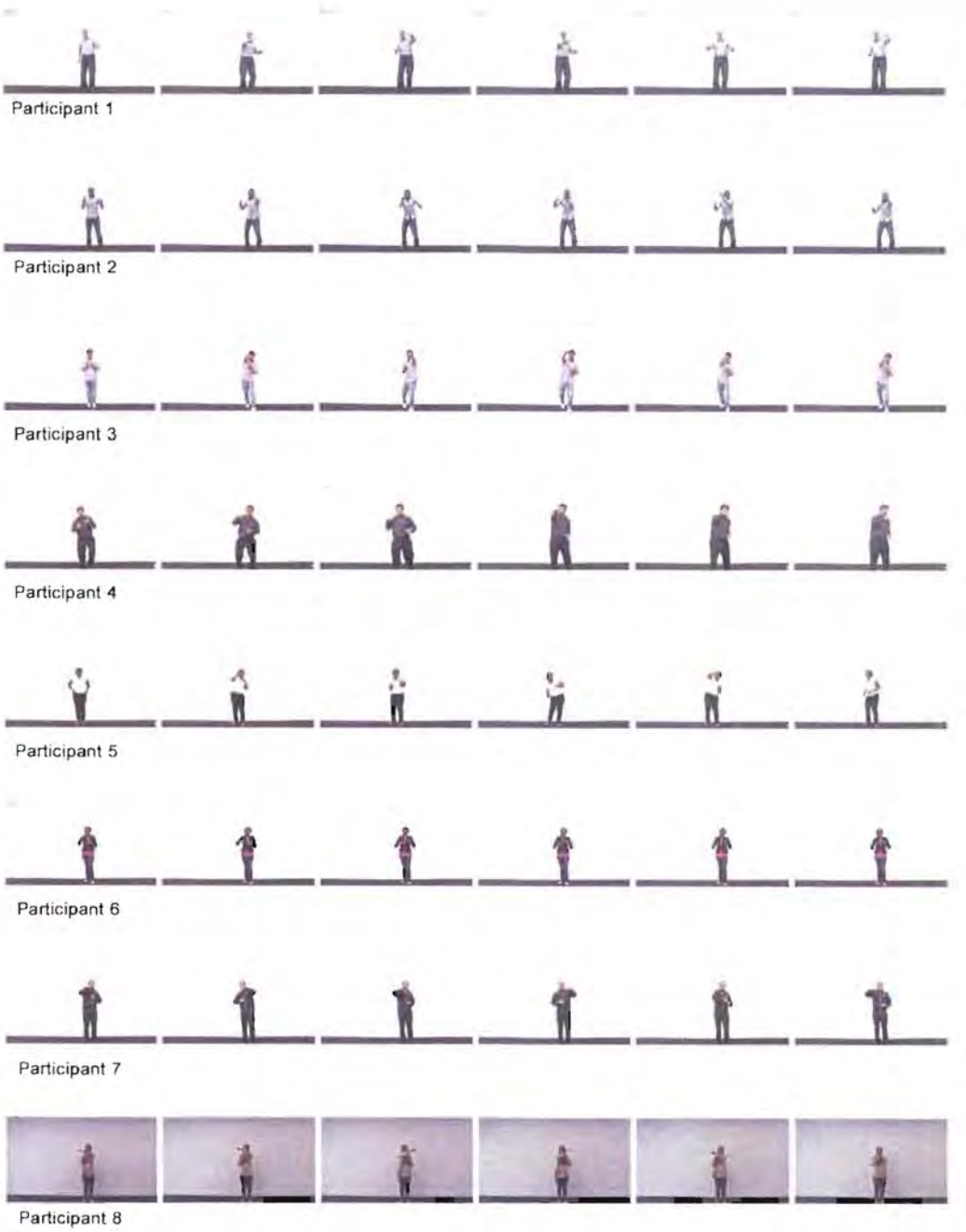


FIGURE 4.8: Subjects performing actions: Boxing

Figure 4.8 exhibits some samples in which participants are performing boxing action.

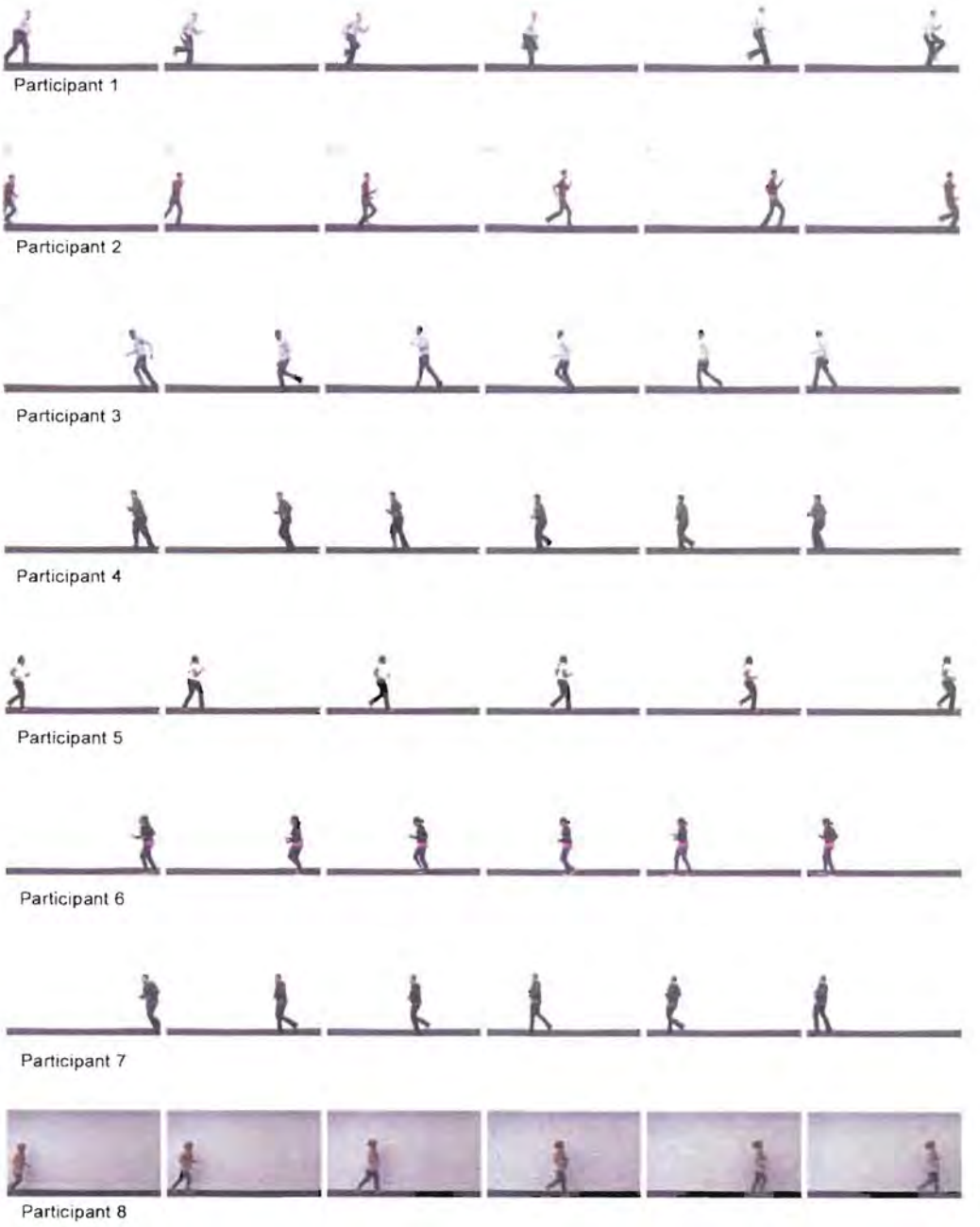


FIGURE 4.9: Subjects performing actions: Running

Figure 4.9 exhibits some samples in which participants are performing running action.

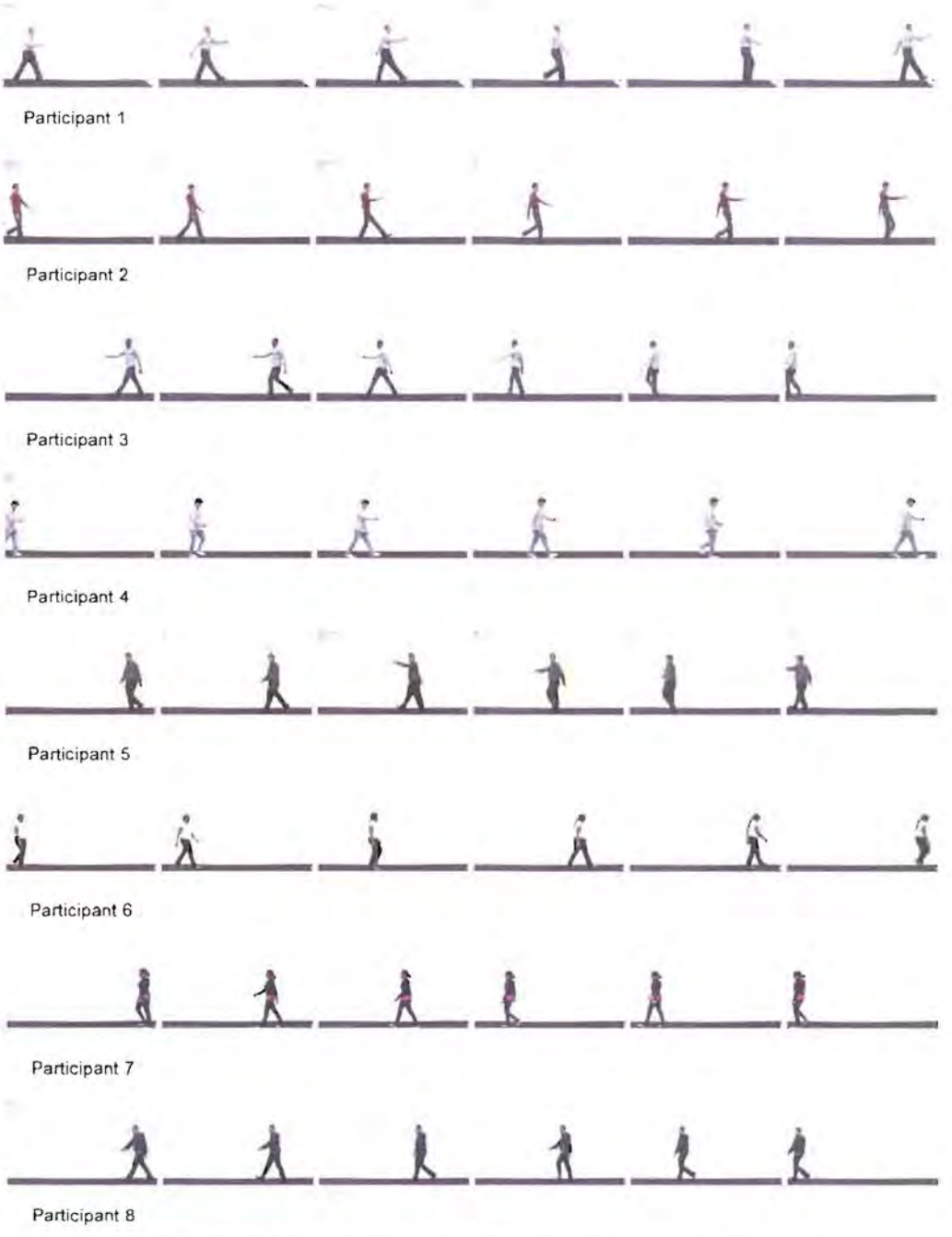


FIGURE 4.10: Subjects performing actions: Walking

Figure 4.10 exhibits some samples in which participants are performing walking action.

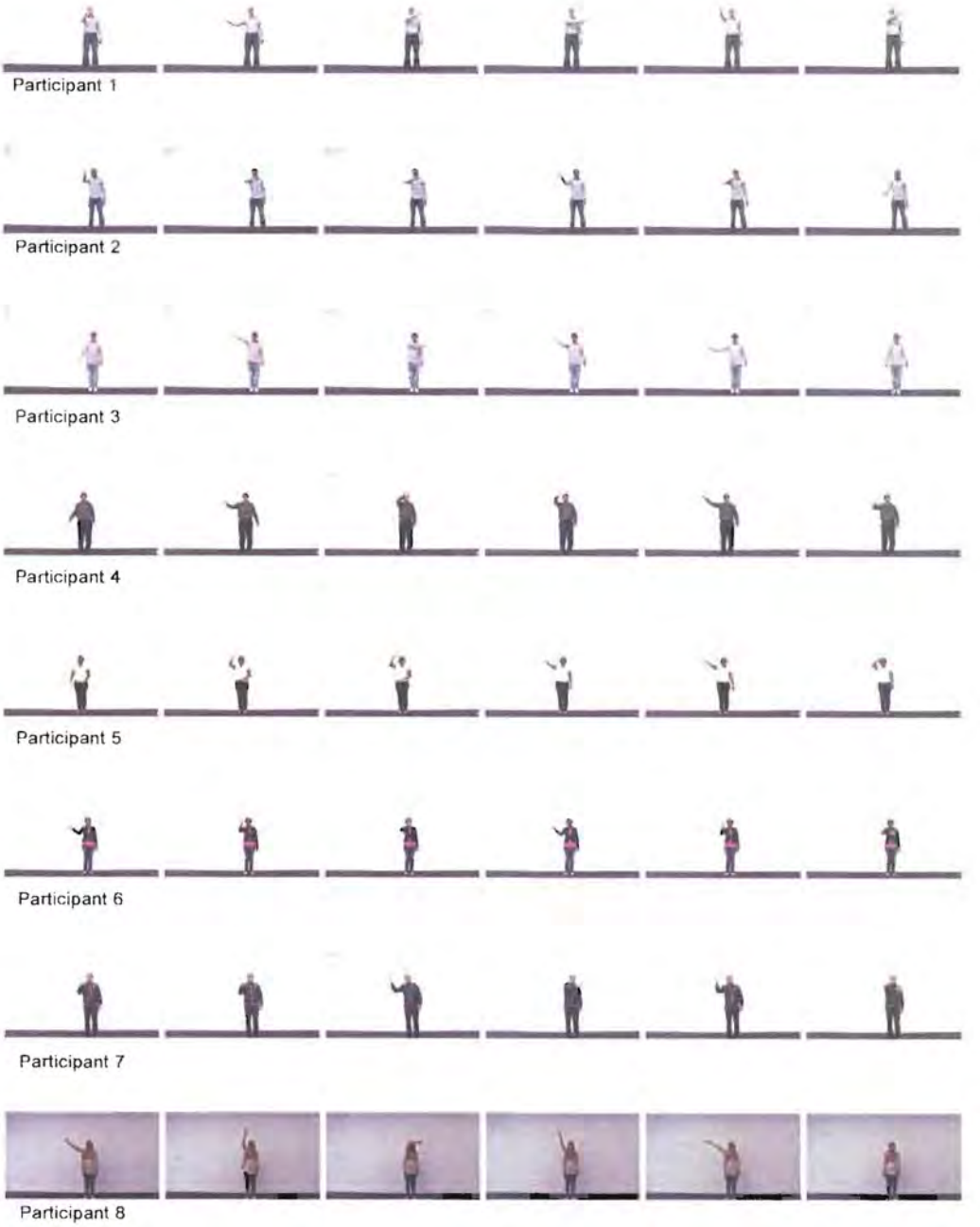


FIGURE 4.11: Subjects performing actions: Waving

Figure 4.11 exhibits some samples in which participants are performing waving action.

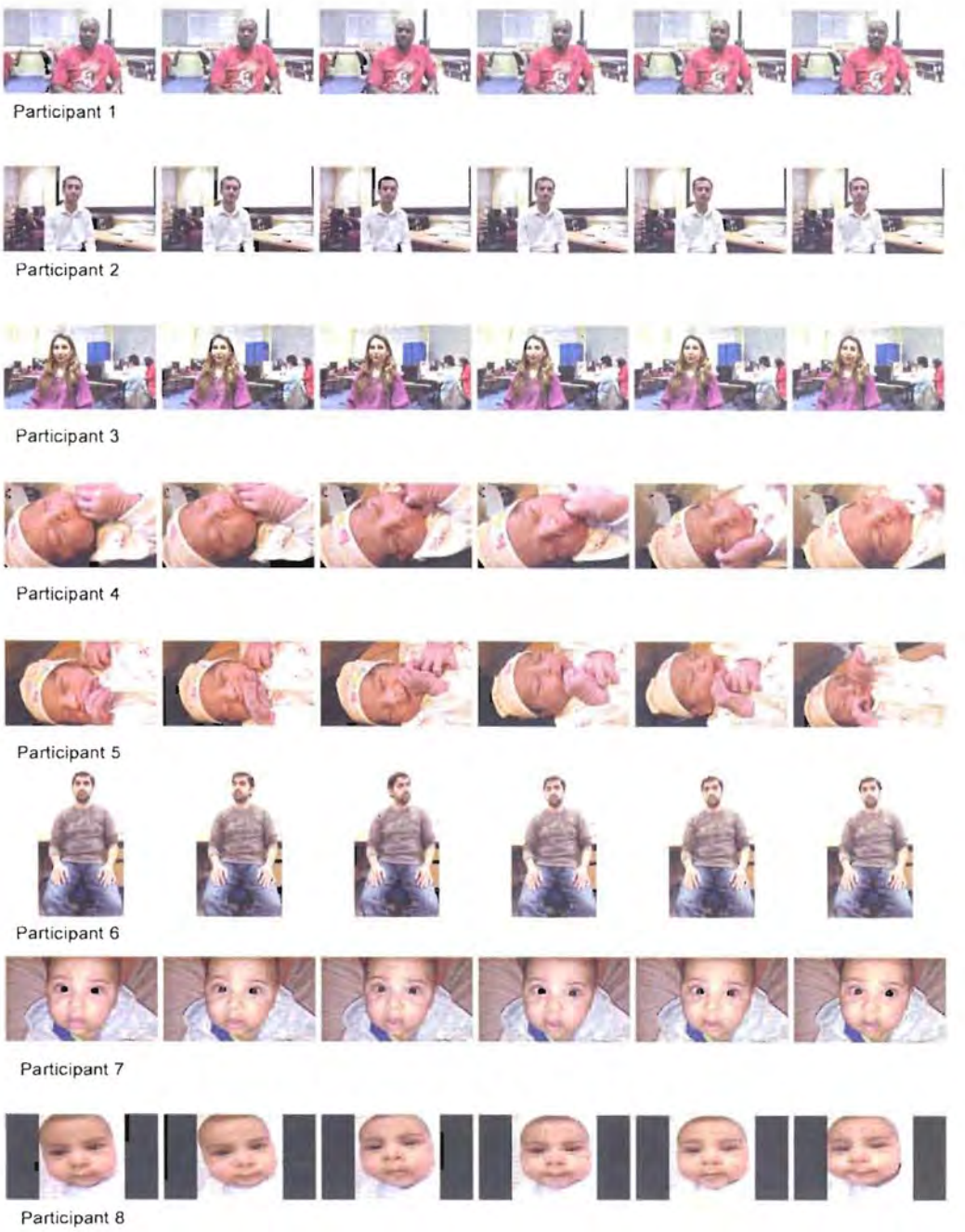


FIGURE 4.12: Close up of subjects

Figure 4.12 exhibits some close up of participants along with some baby participants.

4.5 Summary

In this chapter, corpus generation which is an important step in this research is discussed briefly. We intended to publish this dataset and make it public with the structure as, video ID, video name, segment ID, segment start, segment end, information about age, gender and activity in each segment. The corpus creation is significant for the reasons: a) reparation of data for test and evaluation purposes, b) Restricting the domain of this study to a well-defined and manageable borders, c) Identification of high level features which would be mined by image processing techniques. The foreground is important part in the presence of a human as majority of action is being performed by them. High level features such as actions or gender plays an important role in the understanding of any video. Colour information can also be used to distinguish object and different short boundaries. Corpus analysis is prepared based on best type i.e., user annotation. This corpus provides a base for experimentation and evaluation of the proposed research. Dataset created will be used throughout the process of testing and evaluation in coming chapters. To conclude, the novelty of the framework filed the gap in the area of feature extraction, video indexing and retrieval, which formerly has been treated separately.

Chapter 5

Shot Boundary Detection

5.1 Introduction

Video should be segmented into sections such as scenes, shots and frames before applying indexing, A video sequence can be made up of several shots. These parts are elemental index units in a video database and known as clips. A clip can potentially provide the same functionality in video databases as a word in a text database and to represent them visually key frames are extracted from these clips. Figure 5.1 shows the general hierarchy of video parsing from video stream to frames.

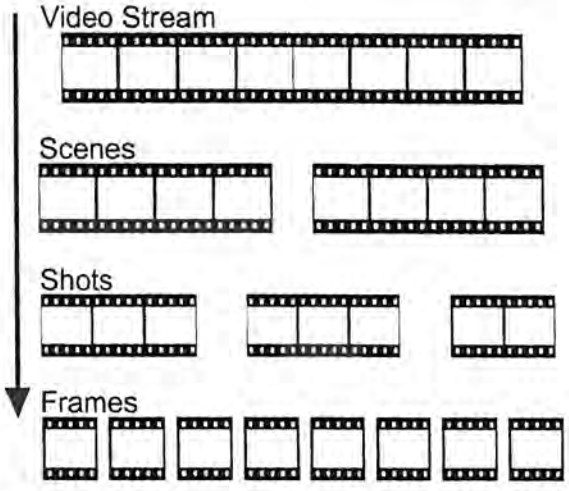


FIGURE 5.1: General hierarchy of video parsing

Zhong, Zhang, and Chang [157] explained in detail that video contents can be characterized by using multimedia features like text, image and speech. Video stream should be spliced into chunks before indexing. Indexing could be performed based on each frame or on whole video. But both these methods are inefficient as if it would be too coarse if performed on whole video and too dense if performed on each frame. That is why researchers commonly indexed on consecutive frames with reference to one theme [158]. Normally video is segmented in terms of shots and scenes.

Hanjalic [159] described a shot as an unbroken sequence of consecutive frames taken by a camera with similar characteristics. There are two types of shot changes, abrupt change and gradual change as described in detail by Yu [160]. Abrupt changes are easy to find and accuracy of 90% is achieved in determining them by using the methods presented in TRECVID [161][162] Gradual transition is classified into three types:

- Fade in/out: Fade in starts with darkness and slowly light comes until full light. Fadeout means start with light and end with darkness

- Dissolve: One scene slowly goes to background (fadeout) by slowly decreasing its intensity and other comes to forefront (fade in)
- Wipe: New scene comes appearing behind a line which moves across the screen

As audio content is synchronized with visual content in video stream, so audio characteristics can be used for shot boundary detection. The relative silence or pause algorithm uses perceptual loudness measure and adaptive threshold is used for classification of non-pause and pause. Pfeiffer [163] proposed algorithm based upon the silence or pause of audio track for segmentation. A shot boundary can be determined by using pause detection in the sound track. Other techniques such as speech recognition can also be used for shot boundary detection. Zhang[108] proposed a technique for shot boundary detection by qualitative change between two frames by counting number of pixels changed from a frame to the next. A shot boundary is found if more than Tb (Threshold value) pixels have changed. Mandal et al. [164] explained that if frame contains color then calculation is done on the basis of RGB (Red, Green, Blue) i.e., all three colors by computing the pixel intensity of adjacent frames and then comparing it with threshold hold. Lee, Yang, and Lee [165] contributed in that by adding that if video is black and white then it can be processed on the value of grey only. This approach is known as Pair-wise pixel comparison and it is not difficult to implement. Shortcoming with this technique is its sensitivity to camera/object motion and noise which results in many pixels change. To overcome this problem histogram can be used for its effectiveness in characterizing the distribution of an image. It statistically shows the intensities of colors in an image. [166]. Histogram technique is more robust against camera movement but still sensitive to lighting condition which results into different graphs for same shot.

5.2 Histogram

The histogram is based on the graphical presentation of the sharing of data. This method was introduced by karl Pearson to generate continuous variables based on the judgement of the possibility of the distribution [167]. The histogram can be represented with a neighbouring quadrangle having tabularized frequencies. The histogram also raised above the discrete intervals, where the observation frequency is equal to the interval area. In addition, the frequencies are divided by the width of the interval named as interval density, which is equal to the height of the quadrangle. Therefore, the numbers are equal to the totalling part of the histogram. The histogram is able to display relative frequencies in a normalized format. Equation 5.1 discusses that number of cases related to the multiple categories along with the part of the histogram. This category is known consecutive and non overlapping for the variables. The chosen intervals are having the same size and neighbouring with each other. To validate the histogram, the variables are in the continuous format because quadrangles are integrated with each other as well. In general, the histogram functionality is based on "mi" which calculates observation numbers. The dislodge category of the histogram is known as "bins". Hence, "x" is the observation numbers and "l" is the accumulative bin numbers. Therefore, histogram satisfies the condition below:

$$x = \sum_{k=1}^{l} h_k \tag{5.1}$$

Moreover, the mapping of the increasing histogram calculates the increasing observations from total number of bins to the specific number of bins. Hence, the representation of the increasing histogram "H_k" is expressed in equation 5.2.

$$H_k = \sum_{m=1}^{n} h_m \tag{5.2}$$

The number of bins is having large and mixed sizes which disclose multiple functionalities of the information. Grant's work is based on 17th century without any guidelines for methodical work in anticipation of Sturge's developed the model in 1926 [168]. The low-density uses bins with wider sizes, and produces a little blur due to the randomness of the samples. The low noise requires high density for the narrow size of the bins. Higher density provides the greatest accuracy to the thickness judgement. Therefore, to vary the width of the bin is beneficial contained by a histogram. Nevertheless, equal width of the bins is extensively used in the system. A number of analytical analysis have discussed and tried to find a better possible figure of bins, but only achieve the detail regarding the outline of the distribution with strong assumptions. The real distribution of the data with multiple objectives of the analysis that provides the bin with diverse widths, which can be appropriated for the multiple experiments required to find the width. In [169] author discusses the variety of effective directives and regulations of thumb rule. The value of the factor of bins "l" can be allocated directly and / or width of the suggested bin "p" can be calculated which is shown in equation 5.3

$$l = \left\lceil \frac{maxi - maxi}{p} \right\rceil \tag{5.3}$$

5.2.1 Histogram Based Image Matching Methods

Stable representation of using color histogram is capable to modify the shape which is largely used for the reorganization of the multiple objects. In [170], Swain and Ballard introduced the techniques related to the indexing colors which resourcefully

recognized histogram colors through the intersection of the histogram algorithm (HI). In the direction of an illumination, insensitive histogram which is based on algorithm having image matching capabilities. There are many alterations with solutions have been suggested, where these methods can be classified into the following groups, where the initial group is based on the following points:

- The Method of the Histrogram Intersection (HI) based on the Conventional System
- The Method of the Matching histogram based on the integrated palette (MPHM)
- The method of the Historgam Inter-section based on Gaussian Weighted (GWHI) [171]

The initial panel had put an effort to design the histogram more robust and resembles to the existing models in order to re-design the new histogram model. In [172], the two authors, Wong and Cheung, suggested the MPHM methods. By using this procedure, the identical colours are more attractable and capture more attractions instead of the perceptual colours. The MPHM has increased the variations of the traditional HI algorithm and also increased its robustness. Proposed GWHI procedure, which designed for the Gaussian weighted functionality to contribute to the matching of multiple colours. This method requires a few seconds for the different dimensions of 40 X 120 to match each other and generate the required results. Further, the GWHI has suggested the other part of the group, which consist of the followings:

- Redesigned the method of GWHI
- (CRG) The Gradients and its ratio of the Colors
- (CECH) The Histogram is based on Color Edge and its Co-occurrence Method

The aims of the other group to produce a histogram and able to extract features from the other colour directly instead of generating other features. In [173], Funt and Finlaysons, discussed the techniques to produce the tri-colours ratio RGB. In addition, in [174], Nayar and Bolles discussed the quantity of the colour and its reflections. Further, in [175], the authors have proposed the CRG for the geometry, shadow, clarification, condition and shadows of the images. At this moment, such procedures have initiated few numbers of spatial data for the appropriate representation of the histogram. On the other hand, the image matching and its colour objects, including the number plates of the vehicles are not to the level of satisfaction. Therefore, to overcome the existing problems, the CECH method is introduced [176]. In [176], only the pairs of pixels that are placed on double edges of the gradient and having the edge points at a certain distance can be the part of the histogram. This proposed design is not only suitable for the special information, but also able to provide the reliability of the histogram and its descriptions.

5.3 Methodology

The fragmentation of video into constituent part is the pre-requisite step for the automatic annotation of any video sequence. Shot cut detection is the fundamental building block for splitting the video into atomic items, which can be annotated and classified. Shot can be classified as a part which is meaningful in its own self. Just like the sentence in a paragraph which gives a complete meaning, shot can be a unit in video, which can convey any distinct information within a video. The discontinuation of temporal information leads to detection to a shot. The features selected for detection of shot change should have the same properties among the images of shots and exhibits clear differences when compared for the shot boundary detection purposes.

Based on extensive literature review, the proposed methodology works on computing the variance metric for frames adjacent to each other and difference is compared to a threshold value. If the difference is in the range of threshold value the frames considered to be of same shot and if difference is out of the threshold value, the shot change can be classified. The flow chart of shot boundary detection technique is outlined in figure 5.2. The process mainly starts with the loading the video file. It extracts the total number of frames from the video sequence. Then the process of detection starts according to the pre-set conditions. The more elaborative illustration can be seen in given pseudo code 1

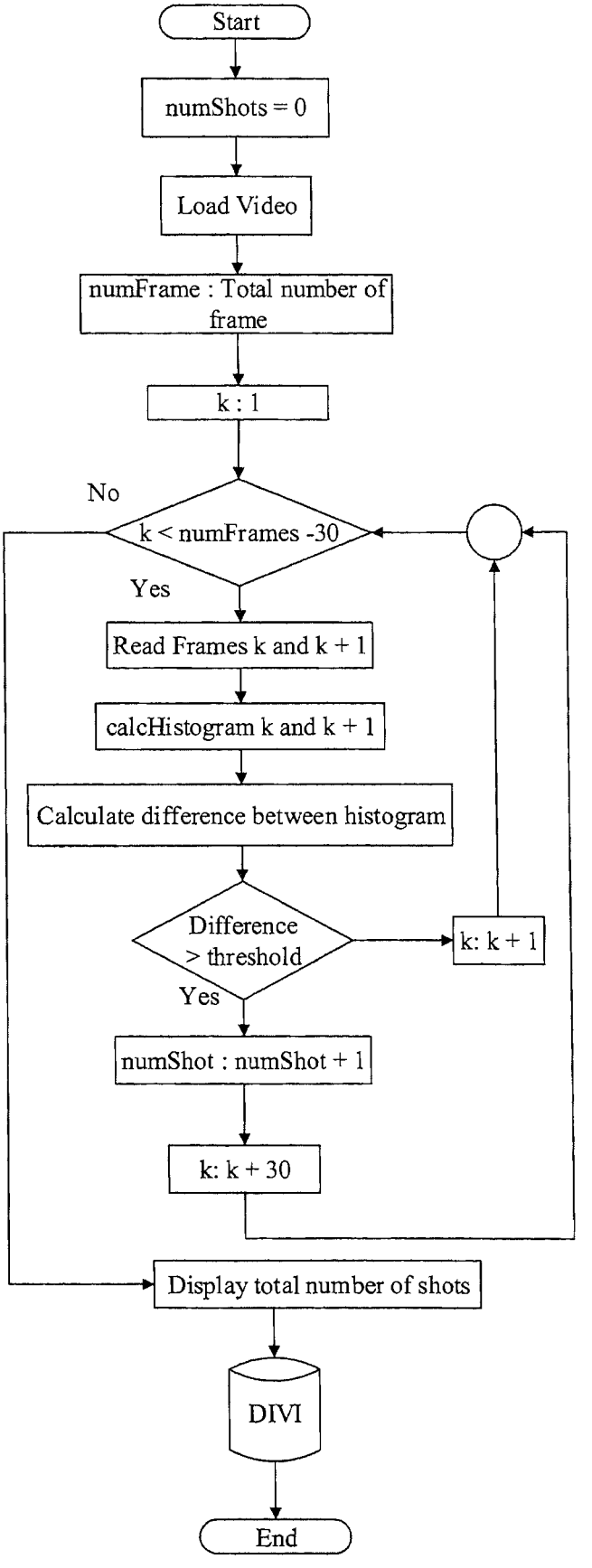


FIGURE 5.2: Flow chart for shot boundary detection

Algorithm 1 Shot Boundary Detection algorithm

```
procedure SHOTBOUNDARYDETECTION(result)                ▷ ShotBoundaryDetection
    video ← ReadVideo
    numOfFrame ← TotalNumberOfFrames
    count ← 0
    threshold ← 0.15                                   ▷ empirically estimated
    numShots ← 0
    while i : 1 . . (numOfFrame − 30) do
        h ← Call procedure Calculate_Histogram for i AND (i + 1)
        diff ← h(i) − h(i + 1)
        if max(diff) ≥ threshold then
            Shot Change Detected
            numShots ← numShots + 1
            i ← i + 30
        else
            i ← i + 1
            No Change in shots
        end if
    end while
    Display Results
end procedure
procedure CALCULATE_HISTOGRAM(histogram)
    img ← frame
    size ← size(img)
    numpx ← size(x) * size(y)
    hr ← redpixelhistogram
    hg ← bluepixelhistogram
    hb ← bluepixelhistogram
    noramlizedH ← (hr + hg + hb)/numpx                 ▷ normalized histogram
    return noramlizedH
end procedure
```

5.4 Results & Evaluation

For evaluation purpose, precision and recall measures were used. Total number is video items are known as collection and by performing different experiments, a definite set of items retrieved from whole collection as depicted in figure 5.3.

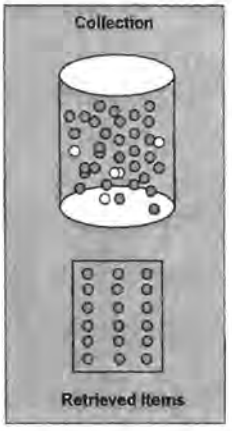


FIGURE 5.3: Collection and retrieved items

The ratio of true positive to true positive and false negative is know as precision as derived in equation 5.4

$$Precision(P) = \frac{tp}{tp + fp} \tag{5.4}$$

The ratio of true positive to true positive and false negative is known as recall as derived in equation 5.5

$$Recall(R) = \frac{tp}{tp + fn} \tag{5.5}$$

Sensitivity or True Positive Ratio is the ratio between true positive and true positive plus false negative as derived in equation 5.6.

$$TPR = \frac{tp}{tp + fn} \tag{5.6}$$

False Positive Rate is the retrieved non-relevant, out of all available non-relevant i.e, ratio of false positive to the sum of false positive and true negative. 5.7

$$FPR = \frac{fp}{tn + fp} \tag{5.7}$$

The harmonic mean of recall and precision is known as F-score which is a single measure of performance of the test. F-score can be formulated as shown in equation 5.8

$$Fscore = 2x\frac{PxR}{P+R} \tag{5.8}$$

The result were evaluated using receiver operation characteristic (ROC). Where by the correct hit or called true positive (TP) is only given for those cases where our algorithm performs correctly. On the other hand, false negative (FN) is equivalent to miss which occurs if the algorithm fails to perform correctly. Moreover, true negative (TN) is corresponding to the undetected areas which are not true (not a wanted object or output). Whereas, those detected areas that do not include the true output are called false positive (FP). However, we are interested in the ratio of the number of true positive out comes to the total number of the existing faces (TP + FN), which is called True Positive Ratio (TPR) or called Hit Rate Recall Sensitivity. In addition to another ratio which is the Positive Predicted Value (PPV), that is represented by the ratio of the number of true positive outcomes (TP) to the total number of the detected regions (TP + FP). This ratio is also called precision value. A generic explanation is given in table 5.1

TABLE 5.1: Precision and Recall

TP	**FN**	**TN**	**FP**
True Positive	False Negative	True Negative	False Positive
Valid detected outcome	Valid undetected outcome	Invalid undetected outcome	Invalid detected outcome

5.5 Intraclass Correlation Coefficient

The intraclass correlation refers to the degree of relationship and relatedness between two set of samples and calculation of quantitative consistency among several observers annotating the same system. Ronald Fisher classified the intraclass correlation within analysis of variance (ANOVA) framework , and afterwards this has been categorized in random effects models framework which defines estimators as formulated in equation 5.9 [177].

5.7

$$I_{ij} = \mu + x_j + \epsilon_{ij} \tag{5.9}$$

In equation 5.9 I_{ij} is the group at j^{th} value and observation at i^{th} location. Unobserved effect in gruop j is denoted by x_j which is shared by values, overall unobserved mean is denoted by μ and unobserved noise is denoted by ϵ_{ij}. In identified model, the value of ϵ_{ij} and x_j are expected to be zero for uncorrelated nature among them and both considered to be distributed identically. σ_α^2 denotes the variance of α_j and σ_ϵ^2 denotes the variance of ϵ_{ij} which is formulated in equation 5.10

5.7

$$ICC = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \tag{5.10}$$

In short, it is used to measure the inter-rater or inter-observer variability and agreement. The dataset generated during corpus generation process will be tested using precision, recall measure and inter-observer variability between ground truth (actual user annotation) and true positive values (algorithm detected) using Intraclass Correlation Coefficient (ICC).

The shot boundary detection algorithm as mention in algorithm 1 has been implemented and the screenshots of accurate identification for shot boundary detection is shown in the figure 5.4 and 5.5 from manually generated dataset.

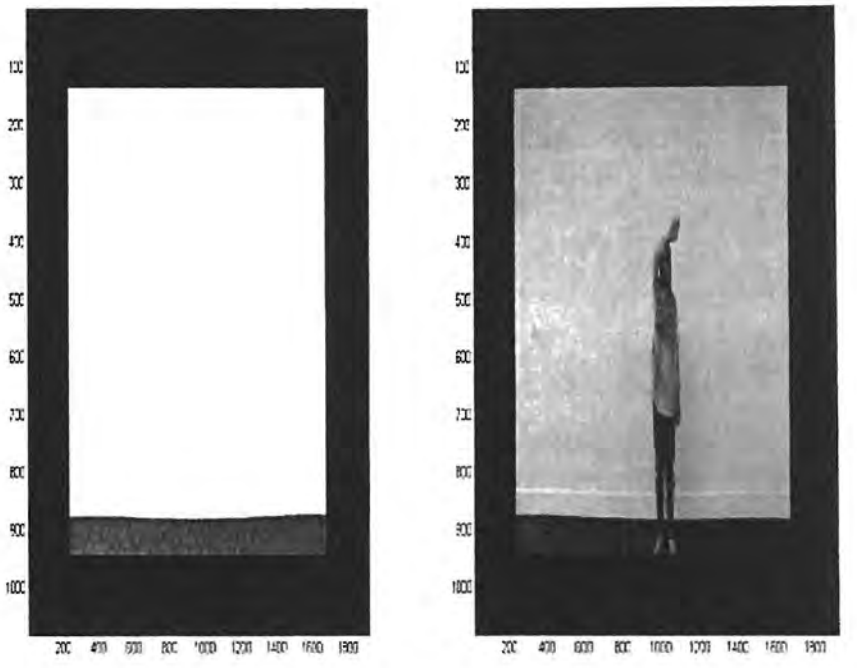


FIGURE 5.4: Shot transition 1

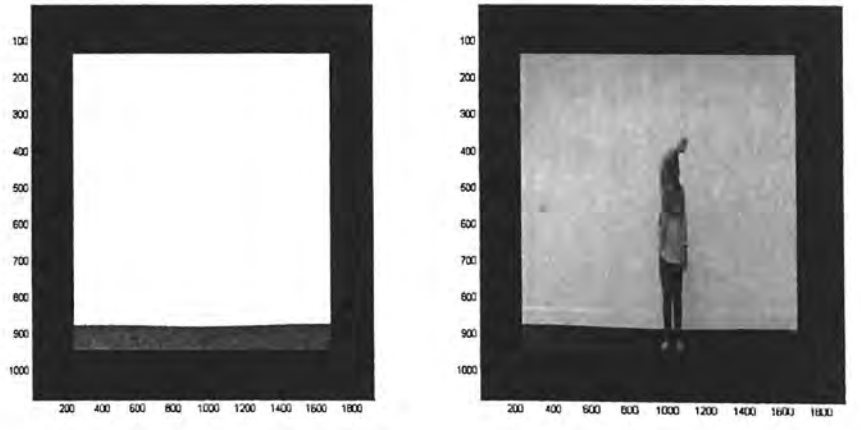


FIGURE 5.5: Shot transition 2

The inter-observer variability calculated with Intraclass Correlation Coefficient (ICC) between actual shots which is actual user annotation values and algorithm detected true positive values comes out to be very good as shown in table 5.2 which shows that there is very less difference between user annotated results and algorithm detected results.

TABLE 5.2: ICC for shot boundary detection

	Intraclass correlation	**95% Confidence Interval**
Single measures	0.7123	0.4479 to 0.8621
Average measures	0.832	0.6187 to 0.9259

The precision and recall along with F measure of shot boundary detection of 25 videos is elaborated in table 5.3. The average precision comes out to be 0.89 and recall is 0.87 which is significant.

TABLE 5.3: Shot boundary detection

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure
1	250	5	4	1	2	1	1	0.78	0.84	0.80
2	250	5	4	1	0	0	0	1.00	0.88	0.93
3	250	5	4	1	1	1	0	0.88	0.90	0.89
4	250	3	3	0	1	1	0	0.80	0.85	0.82
5	250	4	3	1	2	1	1	0.73	0.86	0.79
6	250	6	5	1	2	1	1	0.80	0.89	0.85
7	250	6	5	1	0	0	0	1.00	0.84	0.91
8	250	7	6	1	1	1	0	0.90	0.84	0.86
9	250	6	5	1	1	1	0	0.89	0.88	0.88
10	250	7	6	1	1	1	0	0.90	0.90	0.90
11	250	5	4	1	0	0	0	1.00	0.85	0.92
12	250	7	6	1	2	1	1	0.83	0.86	0.85
13	250	4	4	0	1	1	0	0.85	0.89	0.87
14	250	4	4	0	1	1	0	0.86	0.92	0.89
15	250	6	5	1	1	1	0	0.89	0.88	0.88
16	250	6	5	1	2	1	1	0.81	0.89	0.85
17	250	6	4	2	2	1	1	0.73	0.60	0.66
18	250	6	3	3	1	1	0	0.82	0.50	0.62
19	250	6	5	1	0	0	0	1.00	0.85	0.92
20	250	4	3	1	0	0	0	1.00	0.86	0.92
21	250	4	3	1	2	1	1	0.71	0.85	0.77
22	250	3	3	0	0	0	0	1.00	0.88	0.93
23	250	4	4	0	2	1	1	0.73	0.89	0.80
24	250	7	4	3	1	1	0	0.86	0.60	0.71
25	250	7	4	4	1	1	0	0.80	0.50	0.61
	Average							**0.89**	**0.86**	**0.87**

5.6 Summary

In chapter 5, the shot boundary detection technique is discussed and the procedure has been devised to make use of it for upcoming steps. In the beginning, the introduction of the segmented videos including scenes, shots and frames before applying indexing has been discussed.A shot can be classified as a collection of frames having some common features. Different types of shot transition could occur in a video sequence. Cut detection is the most common one and significant change can be seen in this type of shot change. It also covered the databases who managed the different segments of videos and its features. The histogram and its methods adopted has been dicsussed afterwards. Afterwards, the methodology of the video fragmentation and its sequence which is prerequisite for the proposed framework, flowchart and procedures has been elaborated. Moreover, A shot had classied as a collection of frames having some common features. Different types of shot transition could occur in a video sequence. Cut detection is the most common one and significant change can be seen in this type of shot change. Furthermore, the implementation of cut boundary detection has been discussed. Different classification methods were studied and proposed a framework for cut detection based on feature extraction of histogram and empirical thresholding. The Assumption of no shot transition within one second of previous shot have been made and provided promising results along with better performance.

Chapter 6

Feature detection and analysis

Basic image processing techniques providethe base for detection of any feature within the video. The work starts with implementing the basic image processing techniques for detection of certain feature from the video.These features will be used as building blocks for later tasks. In order to implement any techniques, the video is divided into frames and all the detection operations are performed on the frames i.e., individual image to extract the features from them and to populate the feature set using conventional techniques. Similar to term weights in text retrieval, image features are represented as a vector of real values, which aim to compress high-dimensional image information into a lower dimensional vector space. In the literature, there are mainly three types of image features [178].

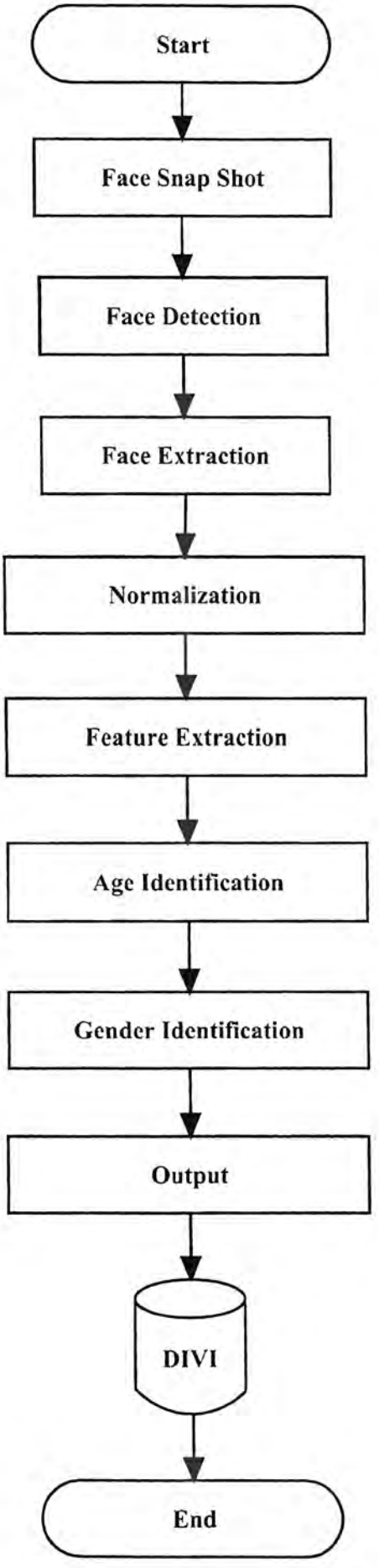
- colour-based features
- texture- based features
- shape-based features

It is not always necessary to construct image features by globally extracting features from the entire image. Although global image features are efficient to compute

and provide reasonable retrieval capabilities, they are very likely to generate unpredictable false positives due to its concise representations [179]. In contrast, image features can be extracted from a finer granularity, such as regular image grids/layouts, automatically segmented image blobs and local feature points. In practice, content-based image classification/retrieval based on regional features usually shows better performance than its counterpart using global features, although it might lead to a higher computational cost in the step of feature extraction. Feature extraction is useful to reduce redundant data, achieved by object classification.

6.1 Proposed Model

The proposed framework starts with the extraction of the facial region from whole frame using face detection technique. Detected face is extracted from the frame and is normalized for further feature extraction. Based on the geometrical and biomechanic feature extracted from the face, age and gender identification is categorized and stored in the database for indexing purpose. The evaluation will be performed using the average Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratios are calculated to get the overall performance of the system. Figure 6.1 shows the steps of the framework.


Start
Face Snap Shot
Face Detection
Face Extraction
Normalization
Feature Extraction
Age Identification
Gender Identification
Output
DIVI
End

As our work is human focussed so the main interest revolves around human. The work starts with the presence of human in the video using the face information. In the next section techniques related with face will be discussed in detail.

6.2 Human Identification

Face plays an important role in identification process. Human presence in a video could be discovered using face identification. Detection of face becomes more challenging due to different factors such as face postures, variability of luminance, different level of contrast and invariant backgrounds. Face detection literature started since early 1970 and evolved by time [180]. Algorithms for face detection can be pigeon-holed into four main categories as below:

1. Identical Template based
2. Appearance based
3. Feature based
4. Facial knowledge based

First two techniques depend on learning and training example sets for concerned objects whereas preceding two methodologies rely on features extracted from facial characteristics and factors manipulation like difference of size, angle and distance. These all will be discussed briefly in the following section.

6.2.1 Identical Template Based

Face can be detected within a set of images by matching a template of it with different images. A template is matched with a set of different images to get the co

relation among images and template to identify whether given set contain any facial information in it. Predefined face template is hovered over the images to see whether it contains any co relation or not. Different researchers used different template for face identification. [181]. To overcome the limitation of angle, distance and size, centroid and angular location are determined and then it is adjusted according to the size of region. A predefined threshold is used to compute the correlation among them. If the threshold value meets the set criteria then image is classified as facial otherwise non-facial.

6.2.2 Appearance Based

In order to check facial feature information, this technique incorporate machine learning techniques and statistical scrutiny. It applies the models which are trained using machine learning techniques. Intensity differences play the key role in training different classifier for machine learning. As classifiers are trained using mono model approaches, this method will perform poor among different variation of images limitations such as occlusion and poor intensity levels. This technique is slower to feature-based but have higher detection percentage. Although this technique is comparatively easy to implement but lacks in efficiency due limitations of adaptability of dealing with shape, scale and pose variations. [182].

6.2.3 Feature Based

Feature based techniques tend to find the face based on actual features of theface which remain the same regardless of changing lighting condition, position and other limitation occurred in face detection. Different researchers used different feature identification techniques to detect face using facial features. By using the Canny edge detector and mean face template detection of face was presented by Phimoltares

et al.[183]. Canny edge detector was used to detect the edges within the images along with theintensity average of similar size images were used to generate face template. It uses advantages of Canny edge detector along with considering the weak edges associated with strong edges. The resultant image finally matched with template to get the final result and if no edges detected then image is discarded. Mondal et al. suggested utilization of geometric shape for detection of human face. Template for center of gravity matching and geometrical features were used to detect human faces. To moderate the noise of image, statistical methods were used such as mean and variance to achieve median filter.

6.2.4 Facial Knowledge Based

In this approach human knowledge is transformed into definite rules. This techniques can also be classified as top down technique. The first step in it is to extract facial features from images. Secondly, identify relationship between different features to represent face contents to encode a set of rules. These rules are helpful in defining the relationship between facial features. Balance nature of rules are defined for accurate detection. Yap *et. al* anticipated that if rules are precise then it can result into failing of detection and miss a lot of faces where as if rules are common then it can result into false detection of many faces [184]. Facial relationship example can be, nose should be relative to eyes and mouth.

Wavelet based identification performs an important role in image processing. The wavelet is based on oscillation waves initializes with the value of zero with an amplitude. The wavelet starts with the value of zero, maximizes and then reduces to zero. The brief oscillation can be visualized classically similar as the monitoring the heart and recorded the seismograph. Usually, the wavelet is crafted for the purpose to obtain the precise values which can be utilized for signal processing. Moreover, the

wavelets can be merged by adopting reverse value with shift, integrate and multiply factors named as convolution. In this technique, the parts of the recognized signal retrieve the information from the unidentified signal.

6.2.5 Wavelet Transforms

A wavelet is based on calculated function which uses divided function for the incremental time signal into the multiple scaling parts. Typically, one can allocate the range of frequency for the each components along its scaling. The waveform transformation is the presentation of the wavelet functionality. The wavelet is translated, copied and scaled is called as daughter wavelet for an infinite size. In addition, the wavelet is rapid decaying oscillated waveform is called mother wavelet. However, there is an advantage of wavelets over the outdated Fourier transforms based the function representation which is having the shrill peaks, discontinuity, non-periodic, deconstruction accuracy and non-stationary signals.

The transformation of the Wavelet is organized into the discrete transformation of the continuous wavelet transforms (DCWTs). It is to be noted that CWT and DWT are having continuous-time (analog) transformation. The wavelet transformation can also be presented in the analog signal where the operation of the CWT can translate to the possible scale along with its translation. The DWT is also the subset of the scale based on the presentation gird and translation values.

A number of techniques to define the wavelet including the wavelet family are discussed in next section.

6.2.5.1 Scaling Filter

One of the most common type of the wavelet is called orthogonal wavelet which is defined by the scaled filter. The scaled filter is named as low pass Finite-Impulse-Response (FIR) for the dimension of 2N and sum of 1. Another type of filter is used for the scaling is called inbo-orthogonal which segregate, decomposes and re-established the filters. To analyse the orthogonal wavelets requires the high pass filter which calculates the quad filter and reorganizes the filters for the reverse timings. The Symlet and Daubechies filters can be described by scaling factors.

6.2.5.2 Scaling Function

The function of the wavelet can be represented by $\Psi(t)$ (i.e. is called the mother of the wavelet) and $\varphi(t)$ is represented the scaling function (also is called the father of the wavelet) as a function of the time domain. The outcome of the wavelet for the band pass filter where the bandwidth can be fragmented of the wavelet. The problem is to organize the complete spectrum for indefinite numbers. The low level of filters are having scaling functions for the transformation is obtained which ensures to cover the infinite numbers. The scaling function of the wavelet is having the compressed state $\varphi(t)$ can be expressed with limited length and is equivalence to the scaling filter.

6.2.6 Human Identification Algorithm

As human face plays an important role in the identification of human presence. Is is clear that a video will contain human if the facial information is present in the frame.If no face is found in the frame then it will be classified as a frame which does not have any human in it.

Haar-wavelet classifier were used for face detection which is the starting point for this application. Haar-like features are used in object recognition using digital image features [185]. Haar-like features are used in object recognition using digital image features. Figure 6.2 depicts standard Haar-like features.

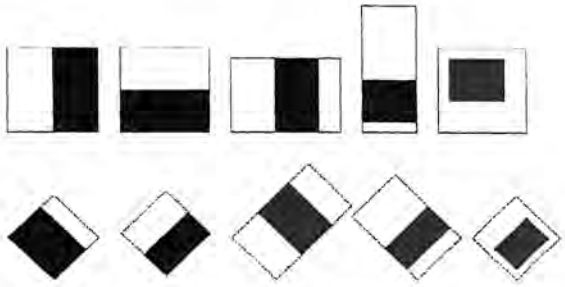


FIGURE 6.2: Haar-like features

Rectangle features shown in figure 6.2 which are relative to the detection enclosing window. We have used these features based on three variation; firstlytwo type of two-rectangle features, secondlyone three-rectangle feature and lastly four-rectangle feature each. Feature's value is calculated as the difference between the sum of the pixels within white and black rectangle regions as derived in equation 6.1 which are used for face detection [185].

$$f_i = \sum(r_i, white) - \sum(r_i, black)$$
$$h_i(x) = \begin{cases} 1 & if \ > threshold \\ -1 & if \ < threshold \end{cases} \tag{6.1}$$

The main advantage of utilizing rectangular features is the performance of them as they compute very fast using technique of integral image.

Integral Image The data structure is based on the integral image, which produces the sum values of the grid in the quadrilateral algorithm for the subset. In 1984, it

was initially presented by Frank Crow to utilize the min-maps. In 2002, the vision of the computer technology in Viola-Jones successfully presented the framework. The principal and study of the possibility distribution function in 2D and 3D is called collective distribution mechanism [185].

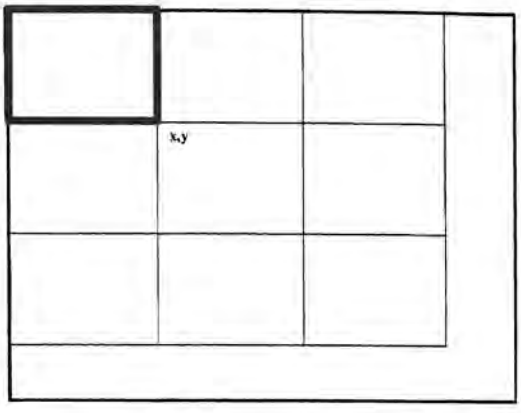


FIGURE 6.3: Integral Image

The rate at any location of x and y in the accumulative area pattern is equal to the additional of all allocated pixels to the top and left of the x as shown in figure 6.3 and its formulation in equation 6.2.

$$I(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y') \tag{6.2}$$

In addition, the area of the table is computed professionally pass only across the image, by the statistic of the accumulated area of x and y described in the equation 6.3

$$I(x,y) = i(x,y) + I(x-1,y) + I(x,y-1) + I(x-1,y-1) \tag{6.3}$$

The quadrilateral and integral image can be accumulated and computed in four arrays as shown in Figure 6.4. The modification between two quadrilateral additions can be evaluated via eight locations. The two quadrilateral is described the features

of the adjacent quadrilateral accumulates of the six array, eight to three quadrilateral locations and three quadrilateral location design the nine to four locations.

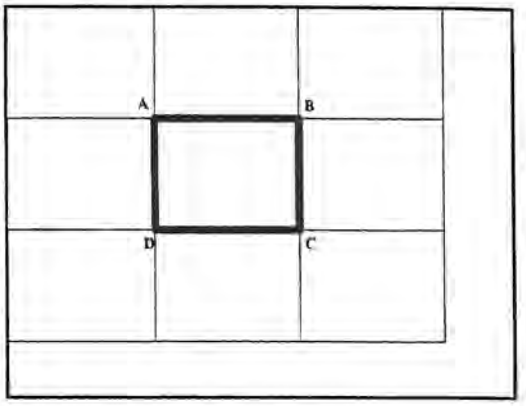


FIGURE 6.4: Rectangular area sum

Once the summed area has been computed to evaluate the task for any quadrilateral in consistent time with four locations of arrays. Especially, identifying the representation in the right hand with the value of A is equal to $x0$ and $y1$, b is equal to $x1$ and $y1$, c is equal to $x1$ and $y0$ and d is equal to $x0$ and $y0$. Therefore, the addition $i(x, y)$ above the quadrilateral spanning by the locations of a, b, c and d. Following is the equation 6.4

$$\sum_{x0 \leq x \leq x1, y0 \leq y \leq y1,} i(x', y') = I(A) + I(C) - I(D) - I(B) \tag{6.4}$$

The dataset generated during corpus generation process is tested using precision, recall measure and inter-observer variability between ground truth (actual user annotation) and true positive values (algorithm detected) using Intraclass Correlation Coefficient (ICC). The inter-observer variability calculated with Intraclass Correlation Coefficient (ICC) between human presence shots which is actual user annotation values and algorithm detected true positive values comes out to be very good

as shown in table 6.1 which clearly shows that there is very small difference between user annotated results and algorithm detected results.

TABLE 6.1: ICC Human presence

	Intraclass correlation	**95% Confidence Interval**
Single measures	0.9653	0.9230 to 0.9846
Average measures	0.9824	0.9600 to 0.9922

Classifiers trained by OpenCV for face detector built on the idea of Haar cascade classifiers. For this work, we used this face detector for the detection of faces. Table 6.2 presents results for the identification of faces among 25 video from the corpus which is developed for this research work. It was a heavily biased dataset where human(s) were present in majority of the video sequence. The table 6.2 shows the Video ID, Total No of Frames, Ground Truth for human presence as annotated by user, True Positive, False Negative, Grouth Truth for non human presence, False Positive, True Negative then the results of Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratio is calculated based on the comparison with the user annotated ground truth. In the end the average average Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratio are calculated to get the overall performance of the system which clearly shows positive results.

TABLE 6.2: Human identification

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure	TPR	FPR
1	250	208	192	16	42	25	17	0.88	0.92	0.90	0.92	0.60
2	250	230	209	21	20	12	8	0.94	0.91	0.93	0.91	0.61
3	250	244	212	32	6	4	2	0.98	0.87	0.92	0.87	0.62
4	250	208	181	27	42	26	16	0.87	0.87	0.87	0.87	0.63
5	250	190	158	32	60	38	22	0.80	0.83	0.82	0.83	0.64
6	250	215	196	19	35	23	12	0.90	0.91	0.90	0.91	0.65
7	250	217	182	35	33	22	11	0.89	0.84	0.87	0.84	0.66
8	250	207	173	34	43	29	14	0.86	0.84	0.85	0.84	0.67
9	250	237	207	30	13	9	4	0.96	0.88	0.92	0.88	0.68
10	250	213	191	22	37	26	11	0.88	0.90	0.89	0.90	0.69
11	250	217	183	34	33	23	10	0.89	0.85	0.87	0.85	0.70
12	250	207	178	29	43	26	17	0.87	0.86	0.87	0.86	0.60
13	250	201	179	22	49	30	19	0.86	0.89	0.87	0.89	0.61
14	250	109	100	9	141	87	54	0.53	0.92	0.68	0.92	0.62
15	250	210	195	15	40	25	15	0.89	0.93	0.91	0.93	0.63
16	250	244	224	20	6	4	2	0.98	0.92	0.95	0.92	0.64
17	250	178	163	15	72	47	25	0.78	0.92	0.84	0.92	0.65
18	250	225	200	25	25	17	9	0.92	0.89	0.91	0.89	0.66
19	250	228	195	33	22	15	7	0.93	0.85	0.89	0.85	0.67
20	250	227	195	32	23	16	7	0.93	0.86	0.89	0.86	0.68
21	250	231	196	35	19	13	6	0.94	0.85	0.89	0.85	0.69
22	250	199	175	24	51	36	15	0.83	0.88	0.85	0.88	0.70
23	250	240	214	26	10	7	3	0.97	0.89	0.93	0.89	0.67
24	250	172	103	69	78	54	24	0.66	0.60	0.63	0.60	0.69
25	250	50	25	25	200	180	20	0.12	0.50	0.20	0.50	0.90
	Average							**0.91**	**0.92**	**0.91**	**0.92**	**0.61**

6.3 Age Recognition

Identifying facial information is a solitary base to validate the age. Several methods to explore facial information have been described in detail by [186][187][188].Kwon and Lobo made the algorithms to estimate facial age for the first time [189]. Two main features i.e. distance and size of different facial characters and amount of wrinkles, were detected using geometrical ratios and deformable contours respectively. Those features have identified face of babies, adults and elderly people. Active Appearance Model (AAM) was another approach used by [190]. It was a coding scheme to assign a face in a lower dimensional room. In this model, age was approximated in the form of quadratic equations which has related the coded representation of faces to the actual age. The results were promising as the specific age estimation was giving improved age estimation when compared to the common aging function of the individuals.

Aging patterns of individuals in a form of datasets were created by [191]. The data sets were consisted of facial images, showing an individual at certain age. In this method, each temporal facial image was judged as a solitary sample that could be presented as a lower dimensional space. The software was substituting an unseen face into different positions in a method that, it could minimise the reconstruction error and pointed out the age of a person as a result. Fu and Huang have described the aging patterns by means of manifold learning on the basis of which, distinguished subspace of learning was created for low-dimensional representations of multiple ages [186]. Regression analysis has described improvements in age estimations using by means of manifold learning.

Most of the literature has described age estimation by taking information from overall face. Suo et al. has developed an alternative method where he used three-level hierarchical face model to detect the age [192]. Its first level was entire face representation; the second referred to numerous confined facial sections equivalent to diverse

features and; the third level represented the use of wrinkles and hairline information. The results have shown an improved performance to detect age of a subject. Ramanthan and Chellapa have created a different methodology for age identification [193]. Instead of estimating the age, they have assessed the age-difference between pair of faces of a same individual. It was used as a classification task where different vectors between pair of faces were utilised for statistical distribution. The allocation had raised a range of maturation or age, which was used during age-division classification problem.

6.3.1 Age Recognition Algorithm

Age of the person can easily identified using the facial information. The age is been categorized into three group namely, young, old and baby. The process start with image extraction and face recognition from the previous phase. The face extracted from the whole image is resized if its too small or extraordinary large. The process complete process of the system is divided into three phases:

1. location
2. feature extraction
3. classification

Age identification algorithm shown in figure 6.5 was implemented to categorised into three groups.

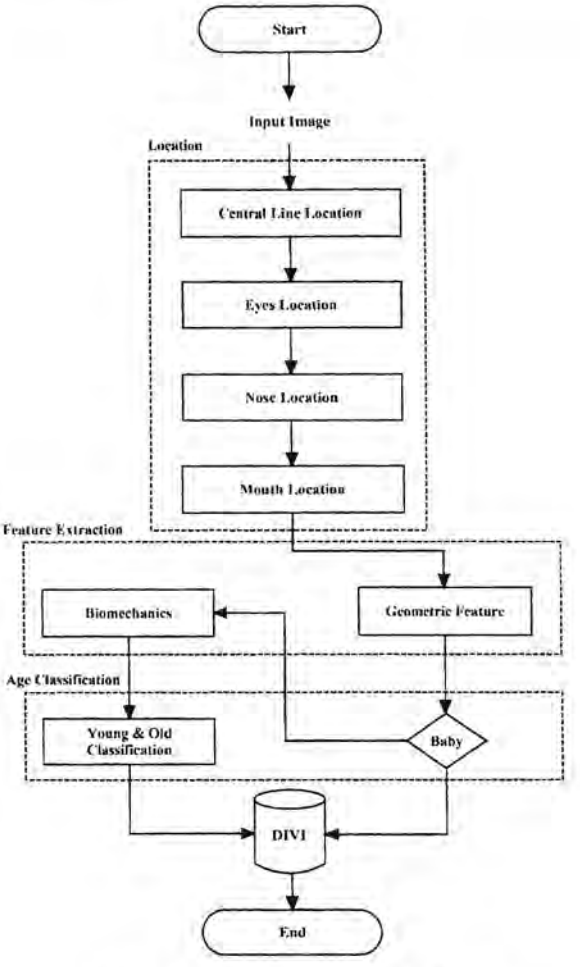


FIGURE 6.5: Algorithm for age

The position of nose, eyes and mouth is located using edge operator. Sobel edge operator and region labelling were used for finding the positions of eyes, nose, and mouth based on the symmetry of human faces and the variation of gray levels. In the location phase, the symmetry of human faces helps to identify vertical central lines of faces. Since eyes, nose, and mouth have significant brightness changes, the Sobel edge operator and region labelling are applied to locate them. Geometric features are used in the system for classification to identify adult and old persons. In the feature extraction phase, two geometric features are evaluated as the ratios of the distances between eyes, nose, and mouth. Lastly, Linde-Buzo-Gray (LBG) algorithm proposed by Lin and Tai was implemented to train the classification system

[194]. The LBG algorithm takes use feature parameters of the vectors.

The dataset generated during corpus generation process is tested using precision, recall measure and inter-observer variability between ground truth (actual user annotation) and true positive values (algorithm detected) using Intraclass Correlation Coefficient (ICC). The inter-observer variability calculated with Intraclass Correlation Coefficient (ICC) between age group (young, baby, old) present in the shots which is actual user annotation values and algorithm detected true positive values comes out to be very good as shown in table 6.3 6.4 and 6.5 which clearly shows that there is very small difference between user annotated results and algorithm detected results.

TABLE 6.3: ICC young person detection

	Intraclass correlation	**95% Confidence Interval**
Single measures	0.9653	0.9230 to 0.9846
Average measures	0.9824	0.9600 to 0.9922

TABLE 6.4: ICC baby detection

	Intraclass correlation	**95% Confidence Interval**
Single measures	0.4703	0.1004 to 0.7260
Average measures	0.6397	0.1825 to 0.8412

TABLE 6.5: ICC old person detection

	Intraclass correlation	**95% Confidence Interval**
Single measures	0.6012	0.2778 to 0.8022
Average measures	0.751	0.4349 to 0.8903

Table 6.6 6.7 and 6.8 presents results for the identification of young, baby and old human among 25 video from the corpus which is developed for this research work. These tables shows the Video ID, Total No of Frames, Ground Truth for human presence as annotated by user, True Positive, False Negative, Grouth Truth for

non human presence, False Positive, True Negative then the results of Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratio is calculated based on the comparison with the user annotated ground truth. In the end the average average Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratio are calculated to get the overall performance of the system which clearly shows positive results. The average precision of young is highest.

TABLE 6.6: Young results

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure	TPR	FPR
1	250	218	174	44	32	19	13	0.90	0.80	0.85	0.80	0.60
2	250	224	181	43	26	16	10	0.92	0.81	0.86	0.81	0.61
3	250	213	175	38	37	23	14	0.88	0.82	0.85	0.82	0.62
4	250	207	172	35	43	27	16	0.86	0.83	0.85	0.83	0.63
5	250	216	181	35	34	22	12	0.89	0.84	0.87	0.84	0.64
6	250	229	195	34	21	14	7	0.93	0.85	0.89	0.85	0.65
7	250	174	150	24	76	50	26	0.75	0.86	0.80	0.86	0.66
8	250	214	186	28	36	24	12	0.89	0.87	0.88	0.87	0.67
9	250	207	182	25	43	29	14	0.86	0.88	0.87	0.88	0.68
10	250	224	199	25	26	18	8	0.92	0.89	0.90	0.89	0.69
11	250	210	168	42	40	28	12	0.86	0.80	0.83	0.80	0.70
12	250	173	140	33	77	46	31	0.75	0.81	0.78	0.81	0.60
13	250	218	179	39	32	20	12	0.90	0.82	0.86	0.82	0.61
14	250	215	178	37	35	22	13	0.89	0.83	0.86	0.83	0.62
15	250	221	186	35	29	18	11	0.91	0.84	0.87	0.84	0.63
16	250	178	151	27	72	46	26	0.77	0.85	0.81	0.85	0.64
17	250	230	198	32	20	13	7	0.94	0.86	0.90	0.86	0.65
18	250	195	170	25	55	36	19	0.82	0.87	0.85	0.87	0.66
19	250	191	168	23	59	40	19	0.81	0.88	0.84	0.88	0.67
20	250	174	155	19	76	52	24	0.75	0.89	0.81	0.89	0.68
21	250	175	140	35	75	52	23	0.73	0.80	0.76	0.80	0.69
22	250	195	158	37	55	39	17	0.80	0.81	0.81	0.81	0.70
23	250	189	155	34	61	41	20	0.79	0.82	0.81	0.82	0.67
24	250	174	144	30	76	52	24	0.73	0.83	0.78	0.83	0.69
25	250	178	107	71	72	63	9	0.63	0.60	0.61	0.60	0.88
	Average							**0.91**	**0.81**	**0.85**	**0.81**	**0.61**

TABLE 6.7: Baby results

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure	TPR	FPR
1	250	206	171	35	44	31	13	0.85	0.83	0.84	0.83	0.70
2	250	207	174	33	43	31	12	0.85	0.84	0.85	0.84	0.71
3	250	199	169	30	51	37	14	0.82	0.85	0.84	0.85	0.72
4	250	172	148	24	78	57	21	0.72	0.86	0.79	0.86	0.73
5	250	236	205	31	14	10	4	0.95	0.87	0.91	0.87	0.74
6	250	189	166	23	61	46	15	0.78	0.88	0.83	0.88	0.75
7	250	171	152	19	79	60	19	0.72	0.89	0.79	0.89	0.76
8	250	188	156	32	62	48	14	0.77	0.83	0.80	0.83	0.77
9	250	238	200	38	12	9	3	0.96	0.84	0.89	0.84	0.78
10	250	211	179	32	39	27	12	0.87	0.85	0.86	0.85	0.70
11	250	187	161	26	63	45	18	0.78	0.86	0.82	0.86	0.71
12	250	179	156	23	71	51	20	0.75	0.87	0.81	0.87	0.72
13	250	209	184	25	41	30	11	0.86	0.88	0.87	0.88	0.73
14	250	179	159	20	71	53	18	0.75	0.89	0.82	0.89	0.74
15	250	242	201	41	8	6	2	0.97	0.83	0.89	0.83	0.75
16	250	185	155	30	65	49	16	0.76	0.84	0.80	0.84	0.76
17	250	189	161	28	61	47	14	0.77	0.85	0.81	0.85	0.77
18	250	221	190	31	29	23	6	0.89	0.86	0.88	0.86	0.78
19	250	219	191	28	31	22	9	0.90	0.87	0.88	0.87	0.70
20	250	209	184	25	41	29	12	0.86	0.88	0.87	0.88	0.71
21	250	211	188	23	39	28	11	0.87	0.89	0.88	0.89	0.72
22	250	235	206	29	15	11	4	0.95	0.88	0.91	0.88	0.73
23	250	179	160	19	71	53	18	0.75	0.89	0.82	0.89	0.74
24	250	245	147	98	5	4	1	0.98	0.60	0.74	0.60	0.75
25	250	239	120	120	11	10	1	0.92	0.50	0.65	0.50	0.89
	Average							**0.85**	**0.84**	**0.84**	**0.84**	**0.71**

TABLE 6.8: Old results

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure	TPR	FPR
1	250	211	162	49	39	28	11	0.85	0.77	0.81	0.77	0.71
2	250	195	152	43	55	40	15	0.79	0.78	0.79	0.78	0.72
3	250	236	186	50	14	10	4	0.95	0.79	0.86	0.79	0.73
4	250	238	190	48	12	9	3	0.96	0.80	0.87	0.80	0.74
5	250	229	185	44	21	16	5	0.92	0.81	0.86	0.81	0.75
6	250	202	166	36	48	36	12	0.82	0.82	0.82	0.82	0.76
7	250	190	158	32	60	46	14	0.77	0.83	0.80	0.83	0.77
8	250	224	188	36	26	18	8	0.91	0.84	0.87	0.84	0.71
9	250	207	159	48	43	31	12	0.84	0.77	0.80	0.77	0.72
10	250	191	149	42	59	43	16	0.78	0.78	0.78	0.78	0.73
11	250	199	157	42	51	38	13	0.81	0.79	0.80	0.79	0.74
12	250	233	186	47	17	13	4	0.94	0.80	0.86	0.80	0.75
13	250	178	144	34	72	55	17	0.72	0.81	0.77	0.81	0.76
14	250	234	192	42	16	12	4	0.94	0.82	0.88	0.82	0.77
15	250	237	197	40	13	9	4	0.96	0.83	0.89	0.83	0.71
16	250	228	192	36	22	16	6	0.92	0.84	0.88	0.84	0.72
17	250	178	137	41	72	53	19	0.72	0.77	0.75	0.77	0.73
18	250	211	165	46	39	29	10	0.85	0.78	0.81	0.78	0.74
19	250	181	143	38	69	52	17	0.73	0.79	0.76	0.79	0.75
20	250	194	155	39	56	43	13	0.78	0.80	0.79	0.80	0.76
21	250	184	149	35	66	51	15	0.75	0.81	0.78	0.81	0.77
22	250	230	189	41	20	14	6	0.93	0.82	0.87	0.82	0.71
23	250	220	183	37	30	22	8	0.89	0.83	0.86	0.83	0.72
24	250	207	174	33	43	31	12	0.85	0.84	0.84	0.84	0.73
25	250	190	27	163	60	49	11	0.35	0.14	0.20	0.14	0.81
	Average							**0.82**	**0.78**	**0.80**	**0.78**	**0.72**

6.4 Gender Recognition

Gender identification also starts with facial recognition. The video of a person is used as an input to extract essential information for distinguishing male or female. Features like eyebrows, face, chin Adams apple etc helps in its identification. Two main approaches are taken in consideration for gender detection; a) facial features and; b) examining the relation between the facial features. The distance between mouth, nose and eye areas of diverse faces has been utilized for the findings. But, automatic detection and view of the facial regions at different facial positions, is difficult to obtain using these methods.

Low-level information of face image area related to pixels value of image is another famous approach. Among this advancement, the most accepted are histogram of slopes, varied texture features, coefficient of wavelet picture transformation and raw gray-scale pixel values. The classification methods of low-level features do better than methods based on high-level features.

The most primitive effort of using computer vision system to categorize the gender was relied on neural networks. A two-layered fully connected network (SEXNET) was taught by Golomb et al., to detect gender from facial images [195]. Tamura et al. introduced a multi layered neural network to categorize gender from face images of altered motions [196]. Hybrid advancement consisted of set of neural networks and assessment trees, was proposed by Gutta and Wechsler [197]. Abdi et al., introduced a PCA based image representation with perception networks and radial base functions [198]. PCA and neural networks was also exploited and they have reported is good performance as well [199]. SVMs for gender identification were studied by Moghaddam and Yang [200].

16 geometric features e.g. eye brow thickness or pupil to eye brow distance etc, from the frontal facial images was calculated by Poggio and Brunelli to identify the gender [201]. A genetic features subset selection from frontal facial image was utilised

as well [202]. Classification based on multi-model gender method using images and voice was proposed by Walawalker in 2002. Jain et al. has researched on the problem of gender identification using frontal face images. The problem was unravelled using SVM, LDA and ICA classifiers. The method was working for female faces but was not satisfactory for male images due to the presence of moustache, glasses and beard as the focus was generally on geometric features of human face.

6.4.1 Gender Recognition Algorithm

Humans gender can be identified using face information only. Once again it starts with processing on facial parts. The gender detection means to identify the human being as male or female. The system takes the video as input and extracts necessary information and classifies the human being as male or female. Humans gender can be specified easily using face only information. This is the one of the main motto behind this work. Features like males have smaller eyes in proportion to the face than that of eyes of the females are used for classification. As we have human face information, so using characteristics of different facial features like forehead, eyebrows, nose, cheek, top lips length, chin jaw and Adams apple we worked towards gender identification[1]. Figure 6.6 shows some results of face detected from the dataset.

FIGURE 6.6: Results for face detection

[1]http://www.virtualffs.co.uk/index.html

The dataset generated during corpus generation process is tested using precision, recall measure and inter-observer variability between ground truth (actual user annotation) and true positive values (algorithm detected) using Intraclass Correlation Coefficient (ICC). The inter-observer variability calculated with Intraclass Correlation Coefficient (ICC) between gender group (male, female) present in the shots which is actual user annotation values and algorithm detected true positive values comes out to be very good as shown in table 6.9 and 6.10 which clearly shows that there is very small difference between user annotated results and algorithm detected results.

TABLE 6.9: ICC male identification

	Intraclass correlation	95% Confidence Interval
Single measures	0.9071	0.8010 to 0.9580
Average measures	0.9513	0.8895 to 0.9785

TABLE 6.10: ICC female identification

	Intraclass correlation	95% Confidence Interval
Single measures	0.6749	0.3884 to 0.8424
Average measures	0.8059	0.5595 to 0.9145

Table 6.11 and 6.12 presents results for the identification of female and male among 25 video from the corpus which is developed for this research work. These tables shows the Video ID, Total No of Frames, Ground Truth for human presence as annotated by user, True Positive, False Negative, Grouth Truth for non human presence, False Positive, True Negative then the results of Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratio is calculated based on the comparison with the user annotated ground truth. In the end the average average Precision ,Recall, F Measure, True Positive Ratio , False Positive Ratio are calculated to get the overall

performance of the system which clearly shows positive results. The average precision of male is higher as compared to female. The reason for this is the wearing of head scarf and make up.

TABLE 6.11: Female results

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure	TPR	FPR
1	250	178	164	14	72	43	29	0.79	0.92	0.85	0.92	0.60
2	250	130	118	12	120	73	47	0.62	0.91	0.74	0.91	0.61
3	250	144	125	19	106	66	40	0.66	0.87	0.75	0.87	0.62
4	250	155	135	20	95	60	35	0.69	0.87	0.77	0.87	0.63
5	250	190	158	32	60	38	22	0.80	0.83	0.82	0.83	0.64
6	250	185	168	17	65	42	23	0.80	0.91	0.85	0.91	0.65
7	250	197	165	32	53	35	18	0.83	0.84	0.83	0.84	0.66
8	250	207	173	34	43	29	14	0.86	0.84	0.85	0.84	0.67
9	250	237	207	30	13	9	4	0.96	0.88	0.92	0.88	0.68
10	250	213	191	22	37	26	11	0.88	0.90	0.89	0.90	0.69
11	250	217	183	34	33	23	10	0.89	0.85	0.87	0.85	0.70
12	250	207	178	29	43	26	17	0.87	0.86	0.87	0.86	0.60
13	250	201	179	22	49	30	19	0.86	0.89	0.87	0.89	0.61
14	250	109	100	9	141	87	54	0.53	0.92	0.68	0.92	0.62
15	250	110	102	8	140	88	52	0.54	0.93	0.68	0.93	0.63
16	250	144	132	12	106	68	38	0.66	0.92	0.77	0.92	0.64
17	250	178	163	15	72	47	25	0.78	0.92	0.84	0.92	0.65
18	250	185	165	20	65	43	22	0.79	0.89	0.84	0.89	0.66
19	250	188	161	27	62	42	20	0.79	0.85	0.82	0.85	0.67
20	250	173	149	24	77	52	25	0.74	0.86	0.79	0.86	0.68
21	250	181	153	28	69	48	21	0.76	0.85	0.80	0.85	0.69
22	250	199	175	24	51	36	15	0.83	0.88	0.85	0.88	0.70
23	250	185	165	20	65	44	21	0.79	0.89	0.84	0.89	0.67
24	250	172	103	69	78	54	24	0.66	0.60	0.63	0.60	0.69
25	250	127	64	64	123	111	12	0.36	0.50	0.42	0.50	0.90
						Average		**0.70**	**0.92**	**0.79**	**0.92**	**0.61**

TABLE 6.12: Results for male classification

Video ID	Total No of Frames	GT +	True Positive	False Negative	GT -	False Positive	True Negative	Precision	Recall	F Measure	TPR	FPR
1	250	211	165	46	39	26	13	0.86	0.78	0.82	0.78	0.66
2	250	227	179	48	23	15	8	0.92	0.79	0.85	0.79	0.67
3	250	180	144	36	70	48	22	0.75	0.80	0.78	0.80	0.68
4	250	225	182	43	25	17	8	0.91	0.81	0.86	0.81	0.69
5	250	189	155	34	61	43	18	0.78	0.82	0.80	0.82	0.70
6	250	216	179	37	34	24	10	0.88	0.83	0.85	0.83	0.71
7	250	193	162	31	57	41	16	0.80	0.84	0.82	0.84	0.72
8	250	204	173	31	46	34	12	0.84	0.85	0.84	0.85	0.73
9	250	239	206	33	11	8	3	0.96	0.86	0.91	0.86	0.74
10	250	191	166	25	59	44	15	0.79	0.87	0.83	0.87	0.75
11	250	180	158	22	70	46	24	0.77	0.88	0.82	0.88	0.66
12	250	188	167	21	62	42	20	0.80	0.89	0.84	0.89	0.67
13	250	179	140	39	71	48	23	0.74	0.78	0.76	0.78	0.68
14	250	234	185	49	16	11	5	0.94	0.79	0.86	0.79	0.69
15	250	193	154	39	57	40	17	0.79	0.80	0.80	0.80	0.70
16	250	203	164	39	47	33	14	0.83	0.81	0.82	0.81	0.71
17	250	183	150	33	67	48	19	0.76	0.82	0.79	0.82	0.72
18	250	238	198	40	12	9	3	0.96	0.83	0.89	0.83	0.73
19	250	201	169	32	49	36	13	0.82	0.84	0.83	0.84	0.74
20	250	226	192	34	24	18	6	0.91	0.85	0.88	0.85	0.75
21	250	199	171	28	51	34	17	0.84	0.86	0.85	0.86	0.66
22	250	227	197	30	23	15	8	0.93	0.87	0.90	0.87	0.67
23	250	186	164	22	64	44	20	0.79	0.88	0.83	0.88	0.68
24	250	229	204	25	21	14	7	0.93	0.89	0.91	0.89	0.69
25	250	218	109	109	32	29	3	0.79	0.50	0.61	0.50	0.90
	Average							**0.89**	**0.79**	**0.84**	**0.79**	**0.67**

Screenshots of accurate identification for age and gender is shown in the figure 6.7 6.8 and 6.9 from manually generated dataset. Region of interest is extracted from facial image, which is described by the face and removing other extra information which is not a part of the face. This is resized to small scale if its good big for fast processing. Histogram was calculated to find the spread of colours in the image. Then position of facial parts like eyes, nose and mouth was extracted. Based on the biomechanics features image was finally classified as male or female along with their age group from any of three category as young, old or baby.

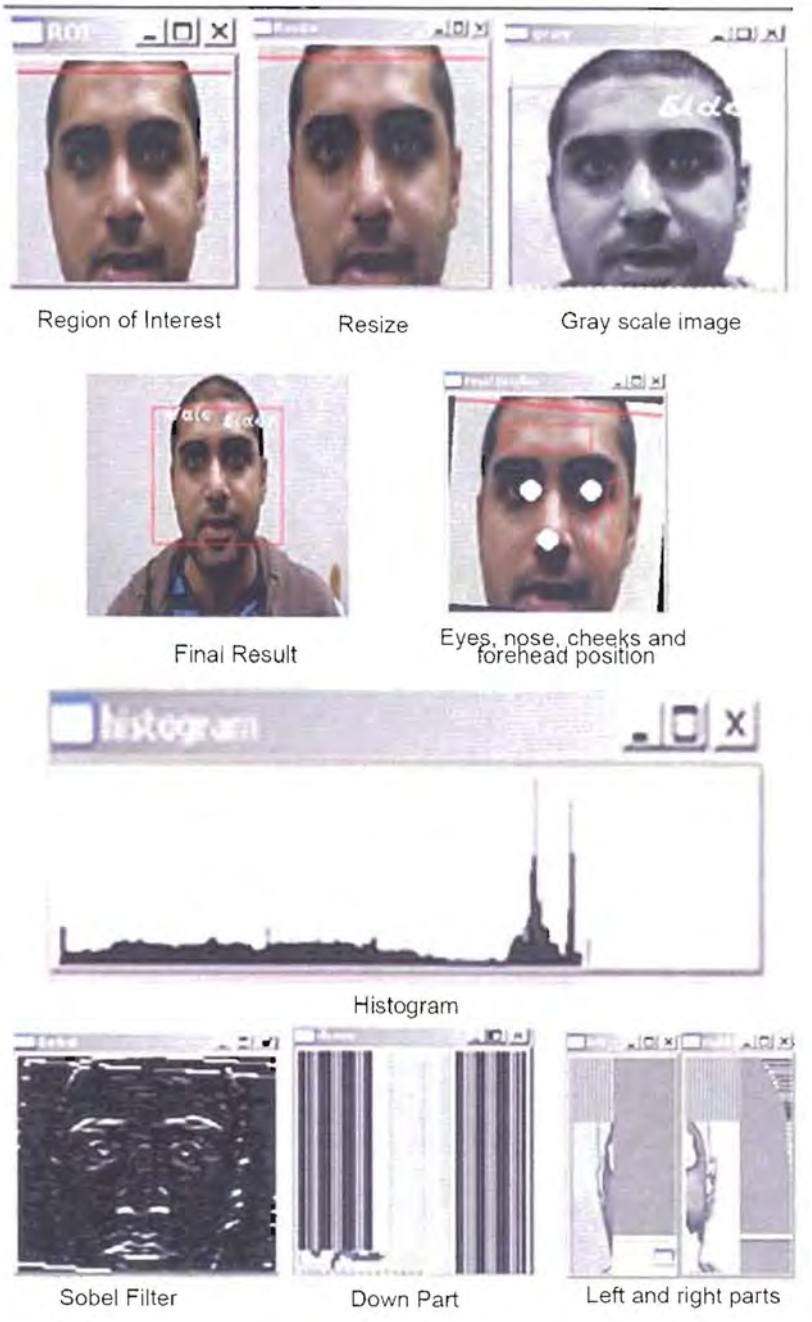


FIGURE 6.7: Correct result for male classification

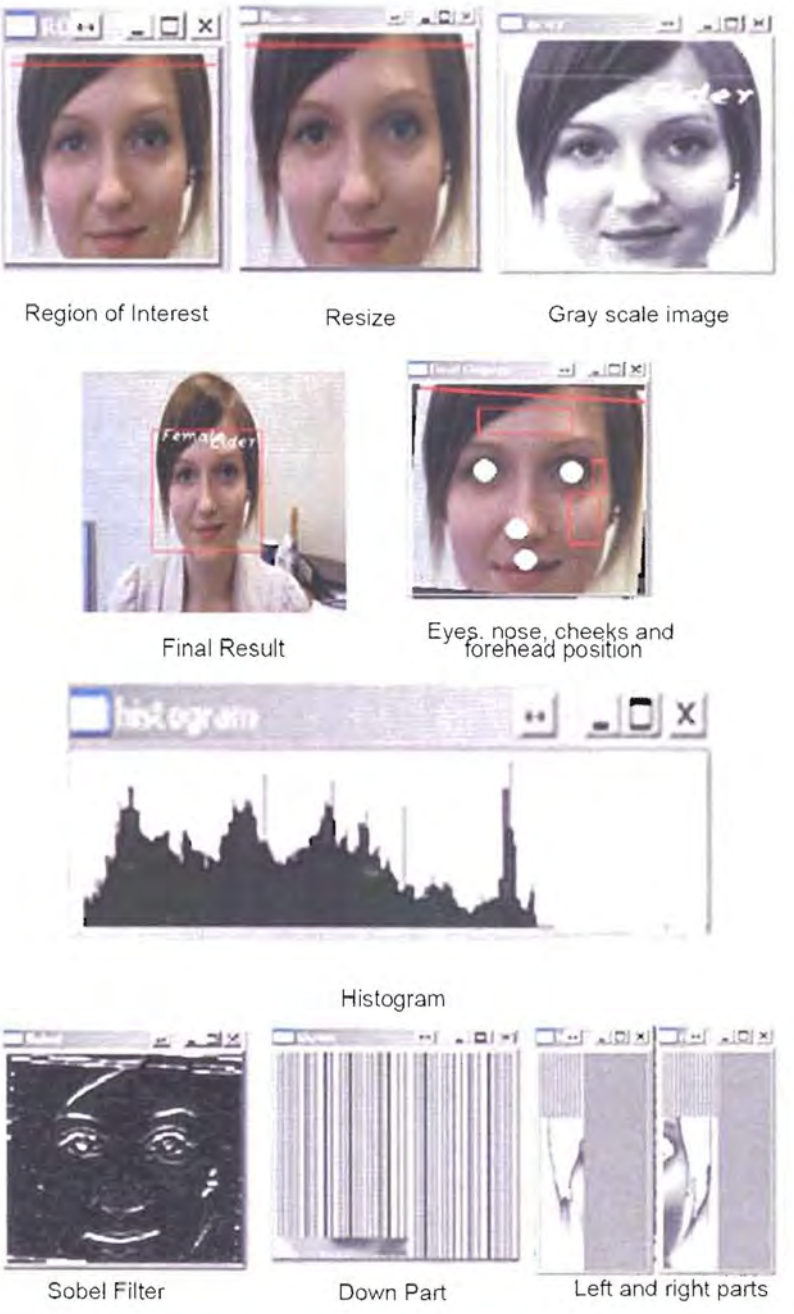


FIGURE 6.8: Correct result for female classification

FIGURE 6.9: Correct result for age and gender

Figure 6.10 depicts some of the unsuccessful detection of age and gender, extracted region of interest, resizing, greyscale conversion, histogram, sobel filter for edge detection, three parts of face as left, right and down, position of nose, eyes, lips, cheeks and forehead and finally classifying wrong as male with reference to age and gender from manually generated dataset.

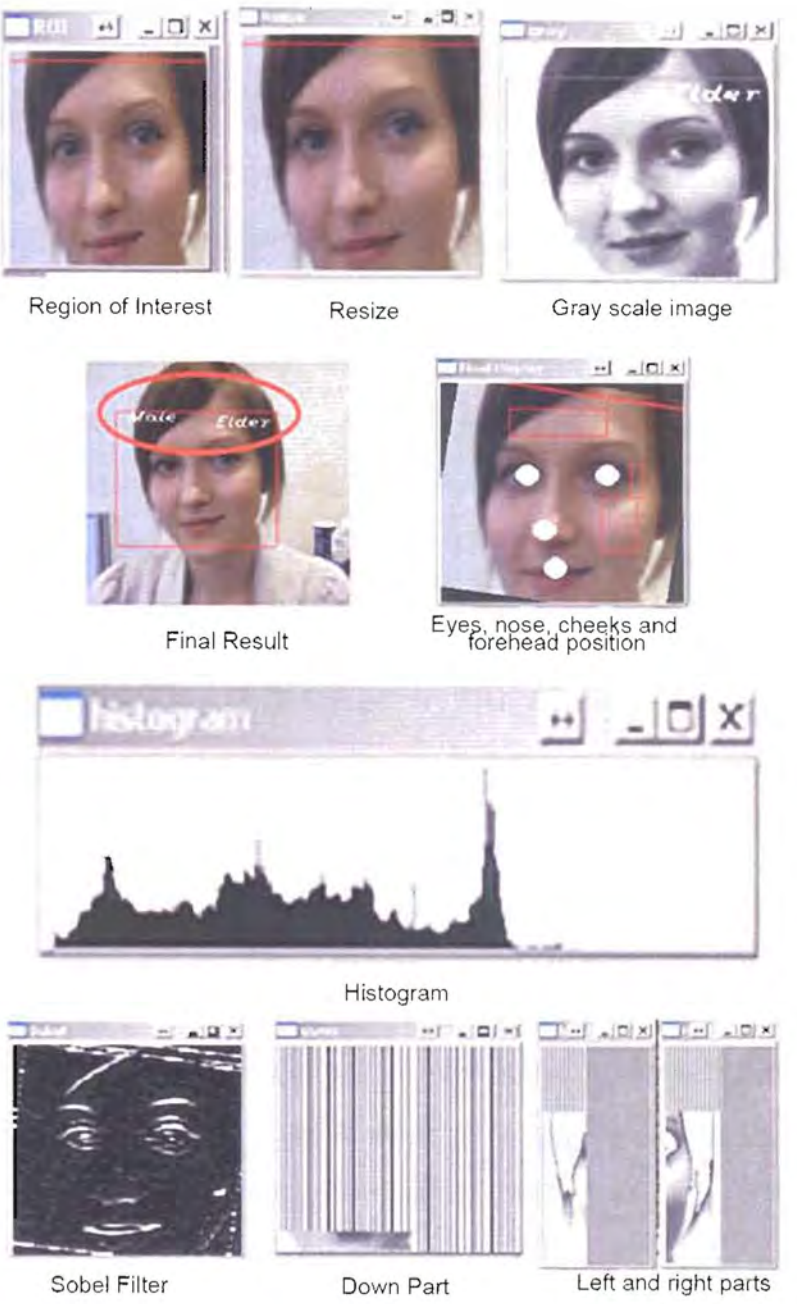


FIGURE 6.10: Incorrect result for female classification

6.5 Activity Detection

There has been a significant increase in digital images and video collection in the last few years due to availability of personalized photo albums, camera videos, feature films and multi-lingual broadcast news in the unstructured or structured format. Now a days the video traffic is around 80% [2] . The data is increased in three different dimensions as video consist of audio, video and textual information. The huge size of data requires qualitative filtering to identify the difference between the relevant and irrelevant information as per the user requirement.

Activity identification helps in the human action recognition. It is observed that the interest lies under explaining complete activity in the video. Depending on the results of single action, a framework can be established to show case output of activity within the video. Identifying simultaneous activities, classifying interleaved activities, uncertainty in interpretation and different residents are few hurdles related to recognition of activity. On the basis of pre-defined activity model, human activities are accurately discovered through human activity recognition. The purpose of activity detection is to determine mutual human activities in real life. It is challenging to accurately determine the human activity because of it being widely spread range and perplex. To construct the activity model different probability based algorithms have been used. The most appropriate Hidden Markov Models (HMM) are being investigated for this task.

HMMs are incorporated for the purpose of action models training. Unique model is trained for distinct action. The body of human is represented in the form of a stick skeleton for activity detection. An action can be transformed into pattern of stick series over the period of time by using these stick sequences which results into sequence of feature vector. Two streams of classification are then further used

[2]http://www.techcrunchies.com/what-percentage-of-internet-traffic-is-video/

to process the feature vector. Weka [3] is explored by one stream classification algorithms. HMM is used by other streams. In order for the HMM to model the action, feature vector sequence must be converted into symbol sequence. Representative stick skeleton of each action type and distance to measure the resemblance between the feature vectors is contained in posture codebook which is designed for this purpose. To explain the training symbol sequence in best manner, the model parameters of HMM are optimized in the training phase. Each feature vector is assigned symbol which is more similar and is matched against the codebook using Estimation of Maximum likelihood parameter EM (Baum Welch). An unidentified sequence can be categorised by calculating the log-likelihood of model sequence with the generated model for unique action. System is implemented for recognition of 4 diverse action types and tested on real human action videos.

6.5.1 Related Review

Action recognition is three step process of identifying the object targeted, generating useful representation and analysing the movement of object. From the above steps, first two develops an object representation which are then applied for image processing techniques and the last two steps are for identification of action pattern and development of semantic description. The recognition process is being facilitated by sequence of images to represent the target object movement. Different recognition can use this basic model for further processing. The movements from an image sequence can be explained by representations. Different recognition methods together with temporal templates, local features and skeletonisation have been investigated. The approaches are summarized as follows.

[3] http://www.cs.waikato.ac.nz/ml/weka/

Local features Histogram is being used largely to represent the local features which points out spatio-temporal features from sequence of images by retrieving the interest points around various events [203] . The important change in both spatial and temporal domain in images is represented by the interest point [204]. The change within the frame is represented by spatial changes, and variation in video sequence is represented by temporal changes. Spatial and temporal scales can be used to capture events. The background is represented by histogram with bright surfaces and major events such as stopping and starting a feet movement by the interest points is represented by dark ellipsoids [203] Gaussian kernel and Harris corner equations can be used to detect interest points by finding the region with major Eigen values. The application of this approach to recognition of human activity has been done successfully. Human action recognition are also using eigenspace methods. In an eigenspace, a curve also called motion curve represents an action given by successive video frames and, by taking a similar action, an unknown action similar to any of memorized motion curves can be judged. High speed human action recognition can be achieved by eigenspace technique but still there is a room for improvement.

Schuldt et al. [156] has used this approach to find descriptors which represent the interest point surrounded by image structure. This can be attained utilizing various combination of optic flow by using histograms, Spatio-temporal gradient, the principal component analysis (PCA) or N-jets (a set of derivatives to the order N). The descriptors when used in collaboration with histogram achieved good performance. Image segmentation or pre-processing is not required but other approaches used special image pre-processing and segmentation.

Skeletonisation In this approach object is represented with the skeleton and transfer a temporal image sequence to a feature vector sequence. Moving light displays (MLD) are 13-18 feature points highlighted on body which forms a 2D stick figure

[205]. 2D labelling which is an example is given by [206] represents a network of connected parts at particular body joints. A new method is anticipated using star skeleton by [207]. This method works by fixing the required pixels in the background and identifies a human object by subtracting the background [208]. The foreground objects is then processed to improve its quality through morphological dilation followed by erosion, this cleans up extracted objects and smoothies their lines. The external outline of the target helps in determining the border points. The five extreme points in the object outline helps in making of star skeleton. These five points represent a head, two hands and two legs. Last but not the least, five points in star fashion with the maximum distances are chosen. This approach is effective in extracting object feature by using a five dimensional vector in two dimensional environment.

Temporal Templates Wren et al., used temporal symbols to develop view- particular images of motions of humans within an image sequence [209]. The foundation is build up on two concepts of where the motion is and how it occurs. For this purpose Motion Energy Image (MEI) AND Motion History Image (MHI) needs to be constructed. The motion of image is represented by MEI which confirms there is a motion activity and shows the angle view. On the contrary the direction of movement is represented by MHI and the brighter pixels shows the most recent movement. To implement this practically, a person should be hold by image sequence and from the other objects available in the image isolate the tracking movement.

Action Recognition The most investigated problem is the modelling of human activates using mathematical model. A training based identification technique Hidden Markov Model (HMM) has been used successfully for speech recognition. This technique is quite helpful and works on transferring the actions into patterns. The HMM technique creates an action by training and is a kind of stochastic state transit model. To identify an action, the HMM with the best action is chosen. Yamato et

al., mentioned that separable features helps in achieving a high identification rate and takes less time to process. Every action has a unique HMM with certain characteristics to monitor human activities. For the first time HMM recognition has been applied [210]. They use three players and six different tennis strokes for HMM. The classification of multi-human activities is done through HMMs by Liu and Chen [211].

Generally a model consists of state transition probabilities, output probabilities and initial state as parameters. To describe the action in best way each category or action is optimized in the learning phase. HMM produces the symbol sequence in the recognition phase. To identify different actions HMM are used together with skeleton representation by [207]. The feature vectors are extracted from images, given a specific symbol to each feature vector by vector quantization (VQ) atlas to store them. To identify the best corresponding action and to train the models symbols are being used. Cunado et al. elucidate the review of approaches for gait recognition[212]. Generally gait identification can be classified as model based analysis and model free analysis. Moving shape and moving motion methods are the two branches of model free analysis. Structural and modular are the forms of model based methods. Different studies use different parameters. Kale *et al.* identified human gait by using HMM [213]. Sundaresan *et al.* calculates the similarity between the feature vectors by using L1 and L2 as different distance metrics for vector difference and normalized inner product of vectors and used binairzed background-subtracted image as feature vector [214].

There are two steps in learning and matching algorithms i.e learning and recognition. Learning is responsible for creating movement from a pool of training data by computing parameters. To find the best match the recognition phase compares the input with each category. Bobick and Davis identify the temporal template by using learning and matching algorithm [215]. Large sample of data is collected to

make algorithm worked and to represent each movement with different views. Using moment based features statistical descriptions are being computed for the temporal templates. Training data are used as descriptors. Mahalanobis distance of input and each movement descriptors was calculated to recognize the test data. The candidate movement object is the movement with similar distance. If the number of candidates is more than one than matching movement will be the one with smallest distance.

Detecting the cyclic motion for specific body movement is one of the many approaches used as action recognition. Autocorrelation and Fourier transformation used legs and torso in this approach. Polana and Nelson designed another approach to recognize repetitive motion represented by features lower level body [216]. The motions are classified using centroid algorithm, to assign the data to particular class achieved by training data which is a kernel based.

6.5.2 Proposed Methodology

A procedure is required to identify single and multiple human actions in video stream to recognize the action. Chen *et. al* stated methods to implement this task [207], even though human body is represented by stick figures [217] in place of star skeleton. From these stick figures low level features are retrieved. Hidden Markov Models were train using these features. The actions of human can be identified through analysing trained HMMs on the basis of most likely performance criterion. To track individual humans a tracking algorithm is applied for multiple humans. Separate, transitive and parallel are three different types of human actions which have been proposed.

6.5.2.1 System Overview

Feature extraction, mapping features to symbols and action recognition using Hidden Markov Models are three different ladders of system architecture as mentioned in figure 6.11. First of all video is chunked into consecutive frames. To extract the feature, background subtraction is being incorporated from present frame and to segment the foreground object, the difference between the current frame and background image is threshold. Once the object is identified, the image is converted into black and white using process of binarisation. Due to binarisation some noise is created and to eliminate the noise from binary image, erosion and dilation along with thresholding image processing methods are used. The posture content from human silhouette is retrieved after the foreground segmentation. Stick figure represents the silhouette. Different features are extracted from this stick figure. In the last phase of feature extraction, the stick skeleton which has been extracted are represented as features vectors for latter action recognition.

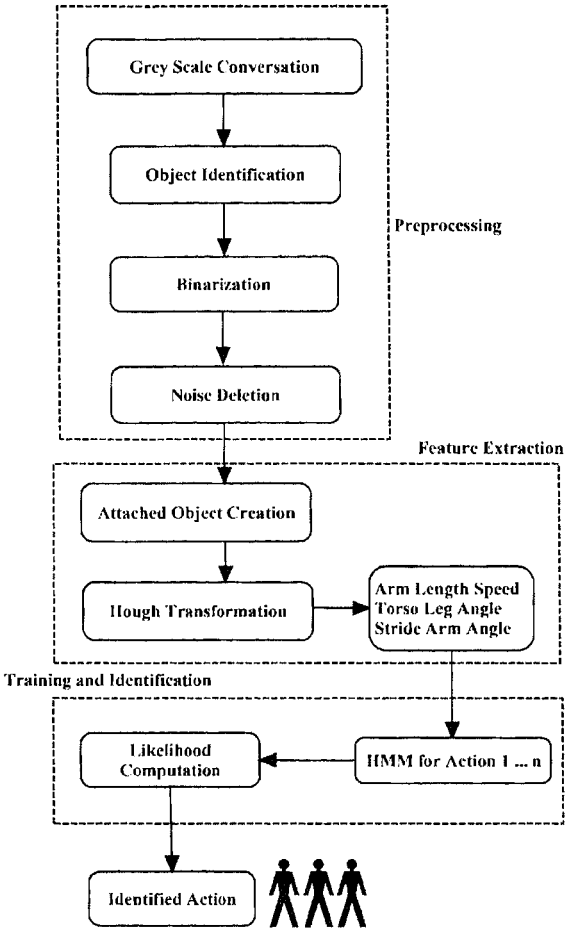


FIGURE 6.11: Action recognition framework

When the background is separated from human i.e object, it is then further processed to find the feature vectors. Human silhouette represents the binary image. Human stick figures are converted from silhouette. From centroid to gross extremities of a human contour the idea of stick skeleton is used. The distance between the centroid and each border point are processed in a clockwise and anti-clockwise direction to measure the gross extremities of human contour. Representative local maximum of distance function are found in extremities. As noise makes it difficult to locate gross extremes. With the use of smoothing filter or low pass filter in the frequency area the distance signal must be smoothed. By locating the finding zero-crossing

of smoothed difference function, local maximum can be detected. By joining these points to target centroid the stick skeleton is established.

The feature from stick figures can't be measured as these are twisted lines. To straighten lines being represented but body parts Hough enhance is applied to stick figure. The important features from the image like arm length, speed, length angle, arm length and torso are calculated.

6.5.2.2 Reprocessing and Feature extraction

The steps for feature extraction are mentioned in figure 6.12. To show intensity information it is converted into grey scale. The object is detected form grey scale image. Precisely, subtraction of background is conceded and the object in the foreground is obtained by applying the threshold difference between background image and current frame image. Resultant image is transformed into binary format for the purpose of dimension reduction. Binarization process fallouts into noise in the image which is detached using the process of dilation and erosion. From this clean image, extraction of the posture contour from the human silhouette is done. As the last phase of feature extraction, a star skeleton technique is applied to describe the posture contour. The extracted star skeletons are denoted as feature vectors for latter action recognition task.

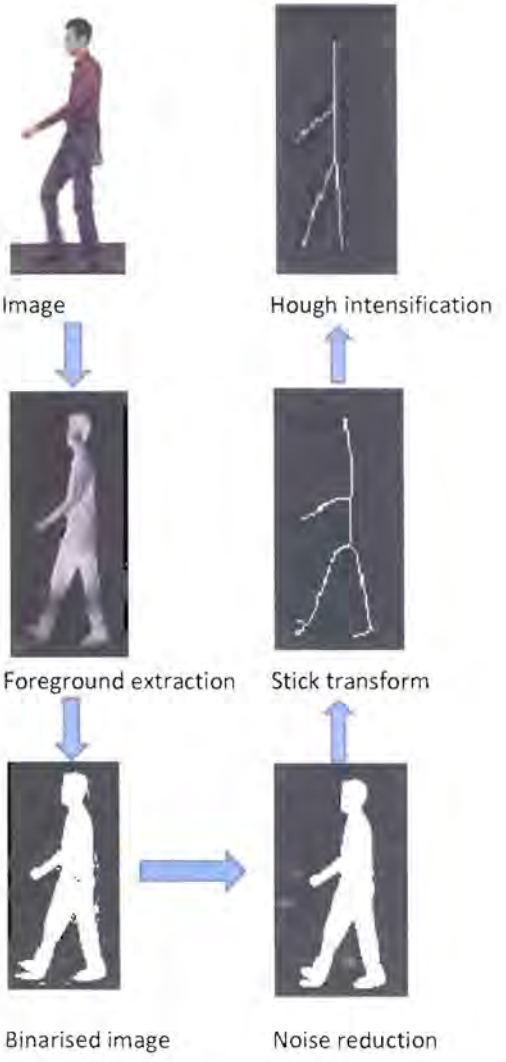


FIGURE 6.12: Steps for feature extraction

The conversion of stick figures from silhouette, thinning operation is applied as mentioned by Staunton [218]. Skeletonisation operation is the concept behind the thinning operation which transforms the silhouette to human to single pixel image also as stick figure. The stick figure is the representation of human body as an arrangement of 7 sticks and 6 joints combining those sticks Cunado et al. [212]. An initial approximation of vertical position of the ankle, knee, pelvis, waist, shoulder and neck for a body of height H set by study of anatomical data to be 0.039H, 0.285, 0.480H, , 0.530H, 0.818 and 0.870H respectively by Drillis and Contini[1] and

this approach is considered as the baseline for human bio-mechanics [219]. A detail annotated skeleton is shown in figure 6.13.

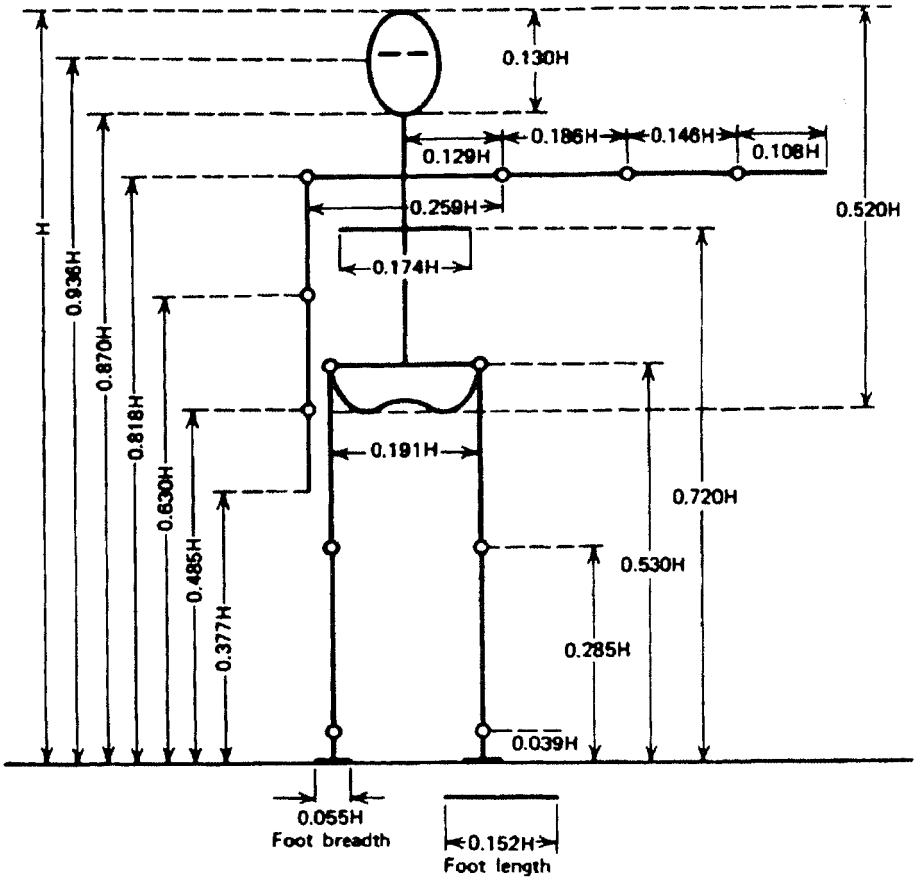


FIGURE 6.13: Length of body segments as a fraction of H height of body [1]

Two border points of each body part p with range constraints can calculate the skeleton. The slope of the lines in linear regression equation can approximate the angles Θ_p of body part p in skeleton data as proposed by Yoo et al. [220] given in equation 6.5.

$$x_p, y_p = [x_i + L_p \cos(\phi + \theta_p) y_i L_p \sin(\phi + \theta_p)] \tag{6.5}$$

Where L_p is the length of body segments, ϕ is the phase, x and y are the coordinates of earlier established position. The combination of these points are 2D stick

figures. The sequence of stick figures gained from silhouette data are the action signature. Leg angle, arm waving speed, arm length, height, torso and arm angle are the significant features used for identification task. Retrieval of stick figures from human silhouette is done through binary image. These features related to action identification are derived from stick figures.

6.5.2.3 Hough Transform

The technique is extracted from the image features and its analysis, vision of the computer system and processing the digital image [221]. The reason of this criteria to get the non-perfect moments of the objects in the class by the procedure of voting. This voting technique is conducted from the students retained from the maxima which is called accumulator designed by the technique of computers formed by Hough transformation. The traditional modification of the Hough transform is related with the identified location of the arbitrary design, called as ellipses and circle. The most useable model now a days is Hough transform is introduced by Richard and Peter in 1972, is also called as generalized Hough transformation [222] based on the patent of Paul Hough [223]. The transformation process is standardized in the computer technology by 1981 journal based on Hough transformation which produce the generalized shapes. During the analysis of numerical images, the partial problems raises the ordinary shapes of the detection such as ellipses, circles and straight lines. In most cases, the corner of the detector can be utilized for the stage of pre-processing to get the pixels over the required edges, but the missing pixels deviates the ideal circle from the noisy points. Due to this reason, the non-trivial groups and its features obtained from the ellipses, lines and circles. The reason of the Hough transformation is to fix this problem by performing the corner points. There is another case of Hough transformation based on linear transformation to detect circle and straight lines. The straight lines can be expressed as $y = mx$,

is called the slope of the line and the value of b is the $y - intercept$. The Hough transformation is a key idea which reflects the physiognomies of the straight line not as the image points of x_1 and y_1, and x_2 and y_2. The slope parameters intercepts the Duda and Hart utilize multiple pair, indicated by r and θ, in the straight lines of the Hough transform. Therefore, these both values are the combination of polar coordinates [222]. This will be explained by taking the example as shown in figure 6.14.

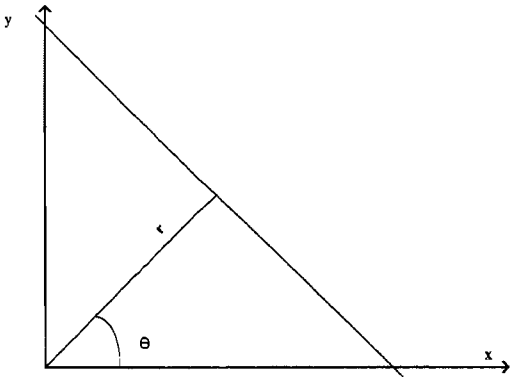


FIGURE 6.14: Hough transform

The value of r indicated the mathematical space between the origin and lines, and the angle of the vector is denoted by θ which refers to the nearest point. The below equation is represented of the line and can be expressed in equation 6.6

$$y = \left(-\frac{\cos\theta}{\sin\theta}\right) x + \left(\frac{r}{\sin\theta}\right) \tag{6.6}$$

Where the re-organized can be arranged for the equation 6.7

$$r = x\cos\theta + y\sin\theta \tag{6.7}$$

It is conceivable to link the lines of the pair of images such as (r, θ) is having an unique values if $\theta \in [0, \pi]$ and $r \in R$, or if $\theta \in [0, 2\pi]$ and $r \geq 0$. The (r, θ) level is directed to the Hough set of the two dimensions of straight lines. This presentation made the transformation of the Hough process is having two dimensional random transformation [5]. The image points with its coordinates can be expressed as, e.g., (x_0, y_0), where the straight lines go through the group of (r, θ) with , So r is the gap among the origin and line is indicated by θ.

$$r(\theta) = x_0 \cos\theta + y_0 \sin\theta \tag{6.8}$$

The sine wave curve indicates the (r, θ) plane, which has distinguish and novel points having the curves corresponds with both locations are imposed to the original line which crosses via both locations. In specific, the group of the points straight lines are able to generate sine wave across the parameters. Therefore, to detect the issues for the co-linear locations must be converted to the issues of getting con-current curves [6].

6.5.2.4 Feature Definition

One procedure to categorise the action is using information associated to motion of skeletal parts. It is important to find out which body part is in motion (e.g., head,legs, hands,etc.). Fujiyoshi et al. cited that walking can be differentiated form running by the angle between the two legs [208]. In the same way the movement of upper limbs helps in recognising actions connected to upper part of human body like waving, boxing and clapping. An assumption is made regarding the feet located on lower extremes of star skeleton. In addition to this human shape and the low pass filter changes according to the number of external points of star skeleton. Particular parts of human body are not necessarily gross extremes. The model of human body

is represented in figure 6.15. Human actions are recognized through structural and dynamic traits. The main structural components of body are torso, arm length, leg length, waving speed and height. Using image processing methods the features of the skeleton are measured. To process it further they are stored as feature vectors.

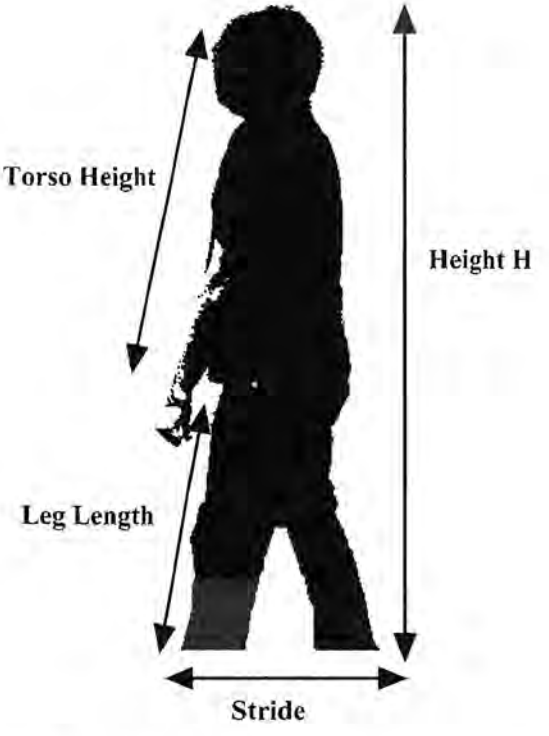


FIGURE 6.15: Human actions trait

6.5.2.5 Vector Quantization

Once the feature is extracted, feature vector are mapped through Vector Quantization (VQ) to symbol sequence. Feature vector of each action are contained in posture codebook and each of them is assigned a symbol codeword. The most identical feature vector in the codebook is the extracted feature vector mapped to symbol. The sequence of posture symbol is the result of mapping features to symbols.

6.5.3 Hidden Markov Models

Let m_{ij} and $n_j(y_t)$ represents probability of evolution from one state s_i to another state s_j. A feature vector of observed state be y_t which is generated the emission probability of sj state. aij refers to transition probability in this case from one stage to another and generated action is represented by $n_j(y_t)$ which is the notation of vector in a feature set. An assumption is made that the feature are not associated among different frames. It aids in managing the calculation for the whole task. The standard discrete HMM implementation and results obtained from them are discussed below. The probability of model is calculated which is giving the feature vectors observed sequence until x_T from x_1. As this is hidden model, L length state sequence is required to generate this sequence. In hidden morkov models the chains are hidden and not known to users. The combined probability produced by model which are producing observations, , may be acquired through the combination of two probabilities i.e. probability of observations and probability of single sequence of state together with adding all potential sequences of state S which is formulated in equation 6.9

$$P(x_1, ..., x_T) = \sum_S P(x_1, ..., x_L, s_1, ..., s_L) = \sum_A P(x_1, ..., x_L | s_1, ..., s_L) P(s_1, ..., s_L) \tag{6.9}$$

The transition probabilities produce results from probability of specific sequence of state which is the main hypothesis of HMM as formulated in equation 6.10.

$$P(s_1, ..., s_T) = m_{Is1} \left(\prod_{l=1}^{L-1} m_{slsl+1} \right) m_{sL} Z \tag{6.10}$$

where m_{slsl+1} is transition probability given time t from l to $l+1$ and $m_{sL}Z$ is the probability until final state of Z from the I initial state. A specific state sequence of time span T can be defined as the product of emission probabilities at specific state which is the hypothesis of feature vector. This could be formulated in equation 6.11.

$$P(x_1, ..., x_L | s_1, ..., s_T) = \prod_{l=1}^{L} n_{sl}(x_l) \tag{6.11}$$

Therefore, the model emitting probability for entire sequence of observation will be formulated as given in equation 6.12

$$P(x_1, ..., x_L) = \sum_{S} m_{Is1} \left(\prod_{l=1}^{L-1} n_{sl}(x_l) m_{sl\ sl+1} \right) n_{sl}(x_L)\ n_{sL}\ Z \tag{6.12}$$

The equation 6.12 can be implemented using algorithm proposed by Holmes and Rubin [224].The features extraction phase will result some features which are employed for probability model estimation. Estimation of likelihood will be used for action recognition after this model is been implemented. Equation 6.13 formulates the calculation of it.

$$\hat{P}(x_1, ..., x_T) = \max\ P(x_1, ..., x_T, s_1, ..., s_L) \tag{6.13}$$

The most likelihood sequence is calculated using Viterbi algorithm [225]. These calculation provides the basics of action recognition and the results of the system will be discussed in next section.

6.5.4 Experiments

In order to analyze the performance of our approach, a system has been implemented to automatically identify four different types of actions. Schuldt et al. provided database which was used for training and recognition of single human action sequences [156]. Six different types of human capital actions are contained in database (boxing, hand clapping, walking, running, boxing and waving) done by 25 subjects. We have used four action related to our approach. These can be divided into two group of actions i.e., boxing and handwaving are the actions related to movement of upper body and walking and running are the actions based on lower body. The database currently consists of 2391 sequences in total. We have extracted the videos according to our proposed technique and the total number of sequences comes out to be 100 in total as each sequence got 25 actors. The homogeneous backgrounds with a static camera of 25fps frame rate are used in all the sequences. The video has a length of four seconds in average and the sequences were down sampled to the 160x120 pixels spatial resolution. This was used as baseline approach and the dataset created during the corpus production will be be tested to be used along with this. Different type of human action in the corpus are shown in figure 6.16

FIGURE 6.16: Different human actions

For identification, every action is calculated using HMM model. The model having highest likelihood is assigned sequence. The graphs represents the analysis of walking in figure 6.17 and hand waveing 6.18. Boxing and waving are recognized easily and categorized appropriately. Handwaving is similar to boxing than walking/running and have second highest likelihood in sequence. Even though running sequence was also categorized correctly but much likelihood is related to walking. It is difficult to distinguish between walking and running and for this reason walking and running have higher likelihood in the graph.

FIGURE 6.17: Most probable human actions: Walking

FIGURE 6.18: Most probable human actions: Hand waveing

All the sequences in the experiment were divided according to the subjects into a training set (8 persons), a test set (9 persons) and a validation set (8 persons). The parameters of each method are optimized through validation and the classifiers were trained on a training set. Table 6.13 explains the confusion matrix of recognition of testing data. The ground truth of action type is on the left and the upper is the recognition action types. The number of each action which are correctly classified

are the number on diagonal. The classification errors are the numbers which are not on the diagonal and the action which system misclassified.

TABLE 6.13: Human actions results

	Walk	**Run**	**Box**	**Wave**
Walk	87.8	4.2	0	0
Run	6.4	80.9	0	0
Box	0	0	82.5	1.1
Wave	0	0	0	89.3

6.6 Summary

This chapter has covered the feature detection and analysis process for human identification , age recognition, gender recognition and activity recognition including in-depth evaluation. Furthermore, the learning, tagging and indexing process has been discussed. Features have been detected on the bases of colour, texture and shape. In this study a model was implemented for human presence using face detection techniques and evaluation was performed using the average precision recall. The algorithms for human identification, face recognition and gender identification, human activity recognition were implemented and evaluated. TP, FN, TN and FP were used to evaluate the sensitivity and specificity. Haar wavelet based human presence gives average precision and recall of 0.91 and 0.92 respectively. HMM based technique used to classify the activities such as walk, wave, box and run. Related activities such as walk and run show the overlapping results that are insignificant. Recognition of males have better precision of 0.89 vs. 0.70 for female while recall in female is 0.92 vs. 0.79 in male. The average precision and the recall for identifying the young are 0.91 and 0.81 respectively followed by 0.85 and 0.84 for the baby and 0.82 and 0.78 for the old. Finally a database is designed to store all these features for retrieval purpose.

Chapter 7

Graphical User Interface & Performance Evaluation

7.1 Introduction

Graphical user interface (GUI) is the key perspective for interacting with any system. It is the main entry point between the computer system and the user. In video retrieval systems, designing the user interface is very important as well. Graphical user interface plays an important role in many perspectives. It is used to interact with the user, displaying require output and for testing the approaches. This is where the actual outcome of the system from the users' perspective could be evaluated/tested. Many-well defined approaches if could not be integrated with the suitable user interface could not serve the purpose. The Human computer interaction (HCI), plays the main role to reduce the gap between the system and the users it and to enhance the interaction of user with systems and to fulfil user needs.

7.2 Human Computer Interaction

In computer science, the interactions between human and computer is an essential essence, and it is also an inherent association in the area of computer science. In general, most of the interaction between computer systems designed for the human in the human context. The study of the Computer Sciences is based on the algorithms, but it can leave certain restraint towards the design of computer system applications. The narrow optimal design is one of the causes to fail the optimization and unable to consider the large number of contextual factors. The general user also known as human user and their circumstances are the main part of the problem in the design, which may not be able to address the complexity ofthe system. In general, the main portion of the programming language in the interactive design is managing the interactions with human. Unprofessional and in sufficient concentration towards users and their task is another cause of the failure among interactions between the computer system and human users. In addition, it also leads towards the system in a risk and may provide the interference to the bad users. The main issue is to consider the combined features of human and contextual part together and compare with the other system which is already established including their designs [226].

Moreover, it is also difficult to justify that how to exceed the rights for the users, even though the system is generating thetrue results. It is unavoidable, but in fact, it shows the progress of the interactive system among the human users. In the last few years, the interactive system introduced the new areas in the computer science, such as forward man, Telecom and industrial engineering, psychology, matching based system and human factors. The human interactions based technology is innovative and useful, but sometimes provides the complex solutions and applications which are difficult for the user's understanding. Therefore, we have to consider the cognitive and social restraint to make successful and robust interactions between computers

and humans. The organizations and researchers conclude that the relation between the human and computer interactions should be organized in a successful way with effective design [227].

At first, the action is based on the implementation of the effective design. However, the interactions of human computer are unable to base on the analysis of usability, but analysis happened not on time. There are few levels of freedom without losing importance and the originality. Xerox Star, Macintosh and Apple cannot be able to do the usability analysis, even though it has an important role in the human-computer interaction system which controls and designs the system. In the second step, the user design based on a centroid system where we must need to consider and emphasize the task based on centered design. The context and key of the system among the people and machine along their interaction of designs is the key allocation. This process is capable to manage and control the machine system operated by humans and resolve their technical problems. In addition, the centered design system is capable to meet the requirements of the task methods and design of the central system. In the third step of the human interaction computer system, an analytical and implementation system merges and generates logical functionality of the centered system. Furthermore, it is capable to structure the core part of the program [228].

The users of the system are not only evaluating the system, but provide the system to the builders who are unable to understand the analytical methods of the system and have weak understanding about the human information system, and its contexts related to the social network. Certainly, computer interactions provide the critical and specialized support forthis system, which has been proven in the area.

Video indexing & retrieval, user interfaces have been used and explored in literature. The major approaches mentioned below are found to be the most commonly used

ones.

1. The informedia interface [229], [230]: Depended upon the video semantic concepts this interface endorsed filtering. When the key word search is taken out than the filters of visual concepts are applicable.

2. The Mediamill interface [231],[232]: This interface amalgamates the concept query visual, text keyword query, and the example query.

The graphical user interface starts with the user requirement gathering which leads to designing and ends with final product testing and evaluation. The overall process of design, development and evaluation will be discussed in thischapter.

7.3 User Interface Designing

In order to test and evaluate the proposed video retrieval approach, the integrated user interface had been designed and developed. The main objective was to satisfy the users' information need i.e. searching the video with the features which are compact and in line with user needs. The main development phase will be discussed in the next section.

7.3.1 Development Phase

Linear Software Development Life Cycle (SDLC), waterfall model has been adopted for development of the user interface for this research. Main phases of waterfall model is shown in figure 7.1 [233].

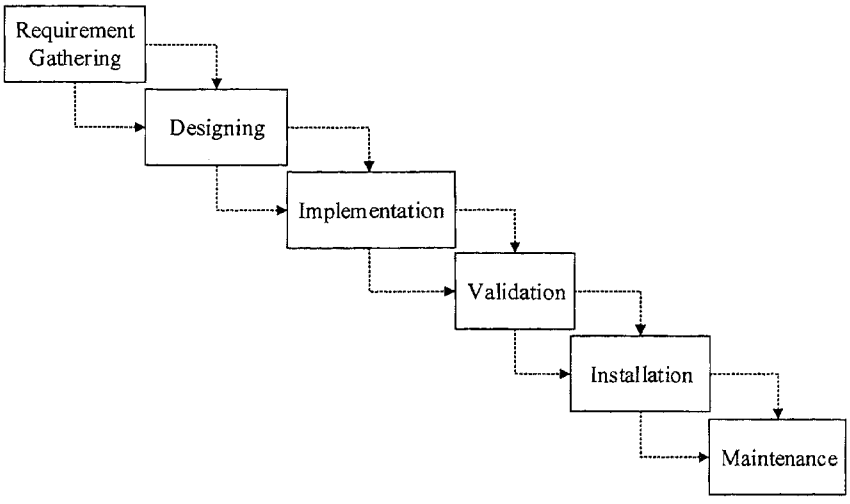


FIGURE 7.1: Waterfall model

As the name suggests, it works from top to bottom just like a water which falls from up to down on a steep. This is the linear model which means it starts with once phase and moves on to next when the phase is finished. The main phases of this model are requirement gathering/analysis, actual designing, implementation/integration, validation/testing, installation/deployment and the maintenance will keep on going until required. Normally the advancements are incorporated before moving on to the next phase once the existing phase is completed. As this model is restricted to one way direction and this type of inflexibility leads to some criticism about it but it is really good for meeting the deadlines. This model is really suitable when the extensive time is spent on requirements and they do not change with the passage of time.

In the first stage, the user needs and requirements have been identified from an extensive literature review. We identify the following major components necessary for video retrieval user interface.

- Search area: query modification option and choices

- Results display area
- Video previewer

Based on the key components of the Video indexing and retrieval architecture, a simplified and user friendly deign has been proposed. User requirements and ease of use along with effectiveness were kept in mind for designing the user interface. Figure 7.2 exhibits the preliminary design of the proposed system [234].

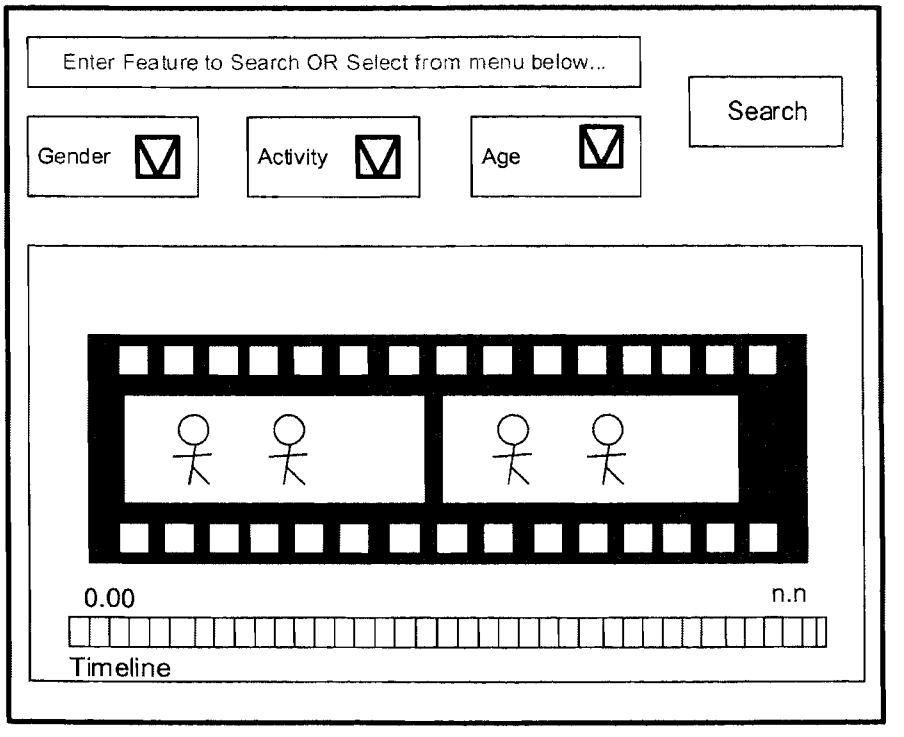


FIGURE 7.2: Graphical User Design

Search area the main entry point of the system. User interests with the system using this area. The searching option provided has two types, i.e. the basic search and the advanced search. The advanced search option provides more control to the user for query refinement. As a user can type in his requirement using a text-based query. The simple search option gives predefined available option for making the query.

User can select from available options. The logical binding among the different aspects of the query can be combined using the radio buttons. For example, the user can select from male, young and walking from the drop-down menu available under gender, age and activity. After selecting the requirement based option, the user can hit enter from the keyboard or user can press the search button using the mouse. Use of logical AND/OR will facilitate the user to modify and refine the search query. The result area comprises of actual results and a video previewer. The result comprises of video name and the shot segments of video, which satisfy the query. The result area also displays the start time of the segment of the video found. These times are click-able and once clicked specified video will start automatically playing from that point. The video previewer shows the actual footage along with indexing marker showing exact location of the query result. The markers are also click-able and display the specific segment once clicked. Once clicked the detail of the running segment is displayed along with the starting time and ending time. User can navigate using either the text based results or the clickable markers available on the video. Figure 7.3 shows the graphical user interface and figure 7.4 elaborate GUI in process.

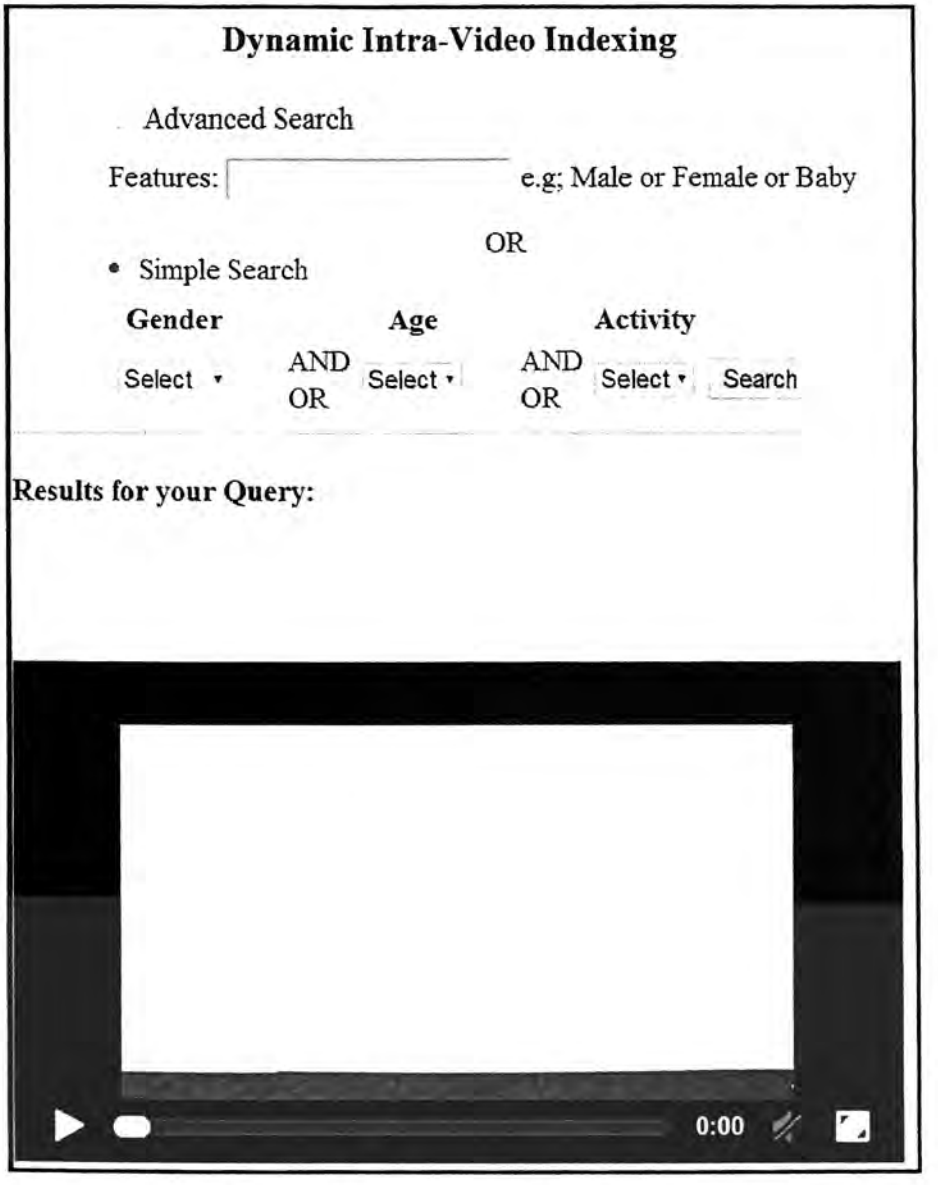


FIGURE 7.3: Graphical User Interface

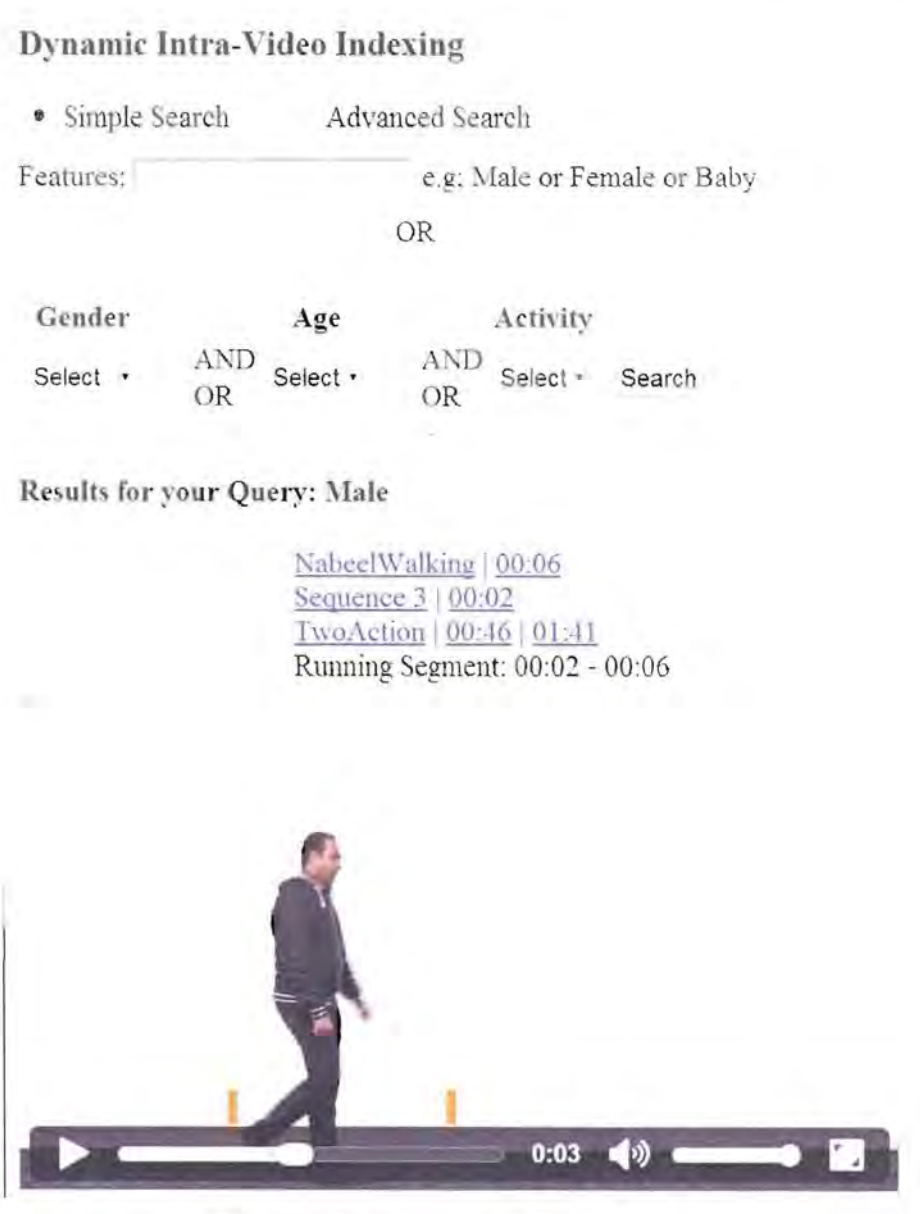


FIGURE 7.4: Graphical User Interface

7.4 User Based Evaluation

The function of usability is used for any technique are used extensively for the evaluation of the system. The main goal of the processing of the usability testing is to test the participant's employee and representative of the spectators who estimate the

usability criteria and meets the requirement. The presence of the illustrative eradicate labelling for the testing such as evaluation of the experts and walk through. In addition, it is not required for the representative's users to be the part of the process. The usability testing handles the tools for the research, and utilize in depth with its investigational methodology. The variety of tests can produce huge number of samples with multiple sizes by using the classical investigations. However, this complex testing only design the information of one participants with qualitative research. Multiple objectives has been achieved by different testing approach with multiple requirements and different timings. The emphasis of literature is presented in simple and unformal way which produces quick result for the industrial and environmental products and its development [235].

The few organizations are having their own point of views for the testing of usability to increase the effectiveness of the products. There are many ways to test the user's benefits such as decreasing and obsoleting the user's frustration and exposure of the issues in representative designs.

7.4.1 Design Evaluation

The key role of the usability is to report the design by obtaining testing information which rectify and identify the drawbacks of the existing material and products before to release in the market. The main purpose of the usability testing is to make sure the improved production of the products as following [236]:

- Is it beneficial and valuable for the potential spectators
- Is it beneficial to understand
- Is it efficient, reliable and effective for the users to meet their requirements

- Is it understandable for the all type of users from the different background to utilize without special training

7.4.2 Elements of Usability Evaluation

This basis elements of usability testing discusses the objectives and research based questions instead of hypothesis:

- The sample of end users can be selected randomly and with criteria
- The environment of the real work on the behalf of representatives
- To consider the end user observation which can be used for the presentation and / or overview of the products and probing of the users
- To gather the performance and important measures for the qualitative and quantitative research
- To recommend the produce with its improved design

7.4.3 Techniques in Usability

The life cycle of product development is applied at different points to satisfy multiple techniques, practices and methods. The consideration of the main methods provide the framework at multiple points in order to support the testing of usability techniques. It is to be noted that the usability techniques are explained in the way where the expansion of the lifecycle would be improved for the products. Following are the research techniques are available which highlight the techniques for building the usability:

- The research of Ethnography
- The selection of the participations
- The research focused group related
- The research relate to the surveying
- The situation of the walk through
- Opening and closing of the card sorting
- The prototype of the research papers
- The evaluation related to the experiential results

7.4.4 Goals of Evaluation

The product design should be arranged in the systematic and organized way with high level of motivations and objectives. The usability is unable to achieve the goals if the proposed design and testing is not convincible. The terminology of the usability should be explained in an organized way for each organizations to achieve the high level goals and required objectives. Following are the key factors which plays an important role to make the product valuable and usable:

- The Efficiency of the Usability
- The effectiveness
- Satisfaction level as per the users requirements
- Accessible
- Robustness

Therefore, the above parameters are having huge contribution to make the required product valuable and usable for the market. Another technical method which makes unique uses of the usability and ensures to meet the user's requirements will be discussed in the next subsection. Figure 7.5 depicts a generalized usability evaluation model.

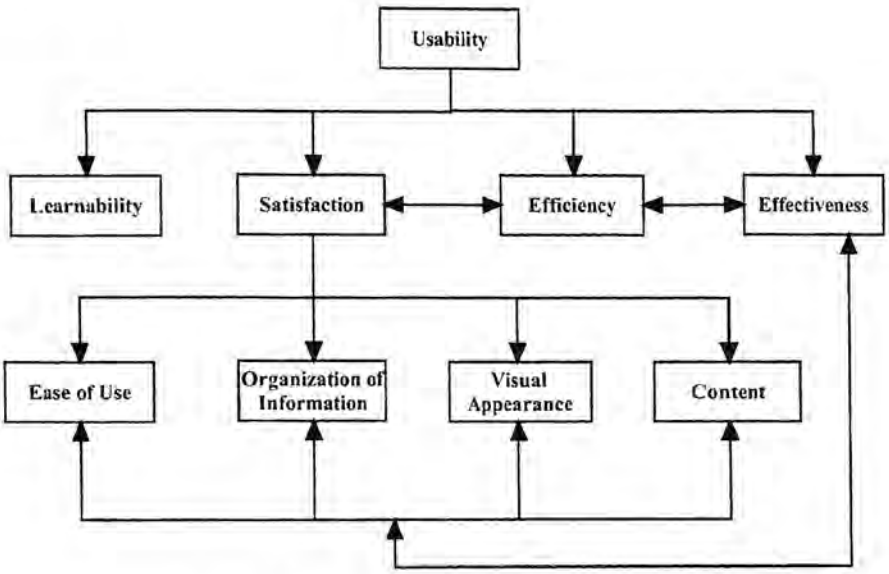


FIGURE 7.5: Usability evaluation model

There are 63 participants in the survey which include both, males and females. The above depicted figure-1 shows that there are around 60% male participants and nearly 40% female participants from the total 63 survey contributors which is shown in figure 7.6.

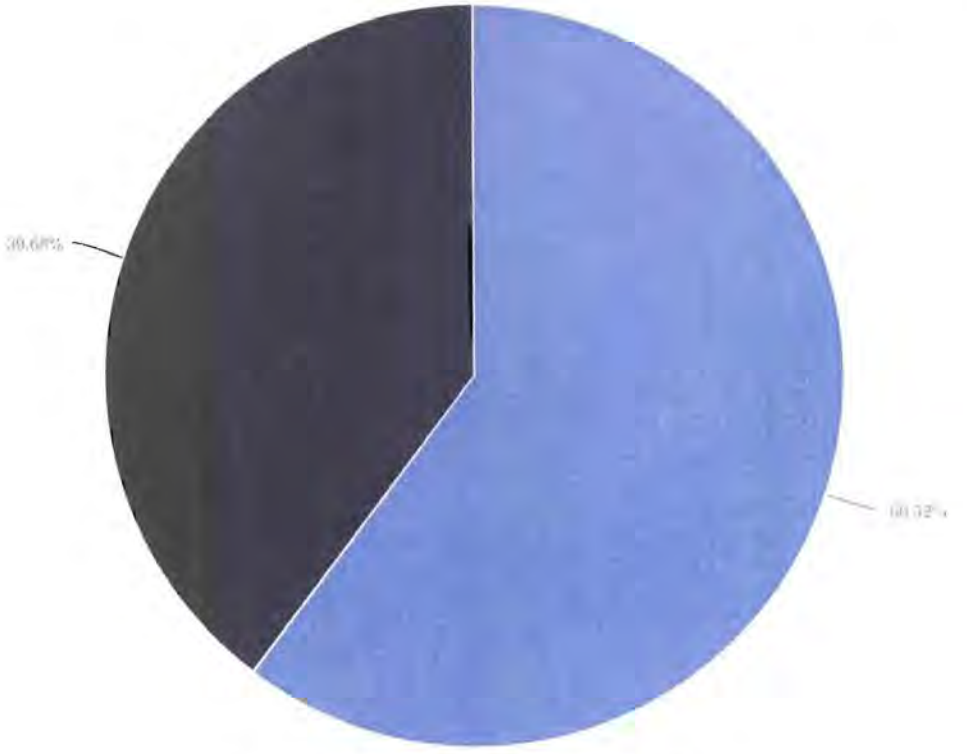


FIGURE 7.6: Male & female participants

The four age groups are developed in the test which are provided to the participants to select one in which their age falls which is depicted in figure 7.7. The mixed responses have been seen in overall survey results where around 46% participants fall under 18-25 age group; this age group has maximum number of respondents. Next age group, 26-39, has more than 41% of respondents and hence has second largest number of responses. Third (40-60) and fourth (above 60) age groups have 12.7% and 0% responses respectively which implies that there is no participant who is above 60 years old; it also infers that there is nearly 1/4th of 18-25 age group participants in the 40-60 age group.

FIGURE 7.7: Age of participants

Then the level of participants' education in which three levels have been set which include 'college', 'undergraduate' and 'postgraduate'. From figure 7.8 It is seen, as a result of survey shown in the Figure-3, that there are around 19% college students, nearly 38% participants are studying in undergraduate level and exactly 42.86% respondents are students at postgraduate level. Highest numbers of responses have been received in the 'postgraduate' level which implies that from 63 survey respondents, majority is studying at postgraduate level.

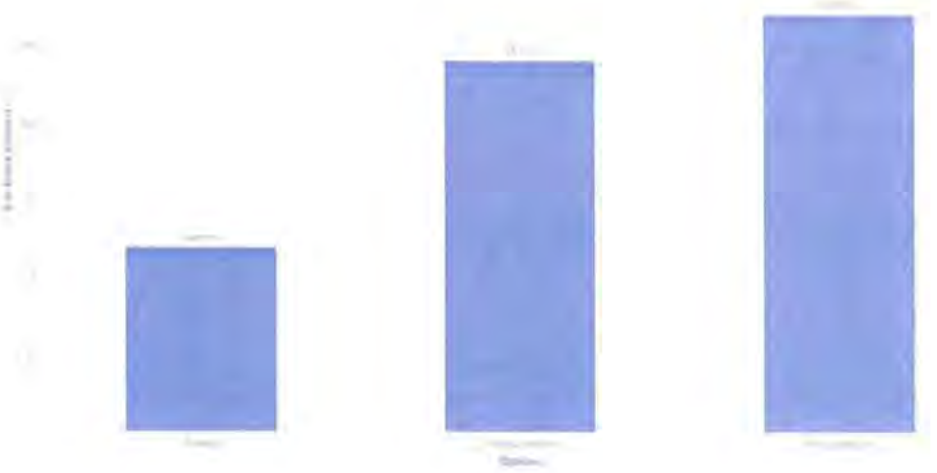


FIGURE 7.8: Level of education of participants

Computer skills of the participants were determined where three levels have been made in the options. It can be seen in the Figure-4 above that 36.51% and 38.1% of survey responses fall in the 'beginner' and 'intermediate' level respectively. However, only 25.4% respondents have 'advance' level of computer skills as shown in figure 7.9. It is implied from the results that majority of the respondents fall under beginners or intermediate levels when it comes to their computer skill.

FIGURE 7.9: Computer skills of participants

Lastly, the visual appearance may different from one platform to another platform which results in different usability results so question is about the use of internet browser through which participants of this survey connects to the World Wide Web (WWW). Three famous internet browsers (Internet Explorer, Mozilla Firefox and Google Chrome) have been set in the options, one of which will be the selection of the participant. Figure 7.10 depicted above shows the results of this question in which it is seen that Internet Explorer users are 22.22%, Mozilla Firefox users are 30.61% and Google Chrome users are 47.62%. From the results, it is inferred that majority of the participants Google Chrome to browse web pages over the internet so the system was optimized for all platforms but more rigid testing performed on Google Chrome browser.

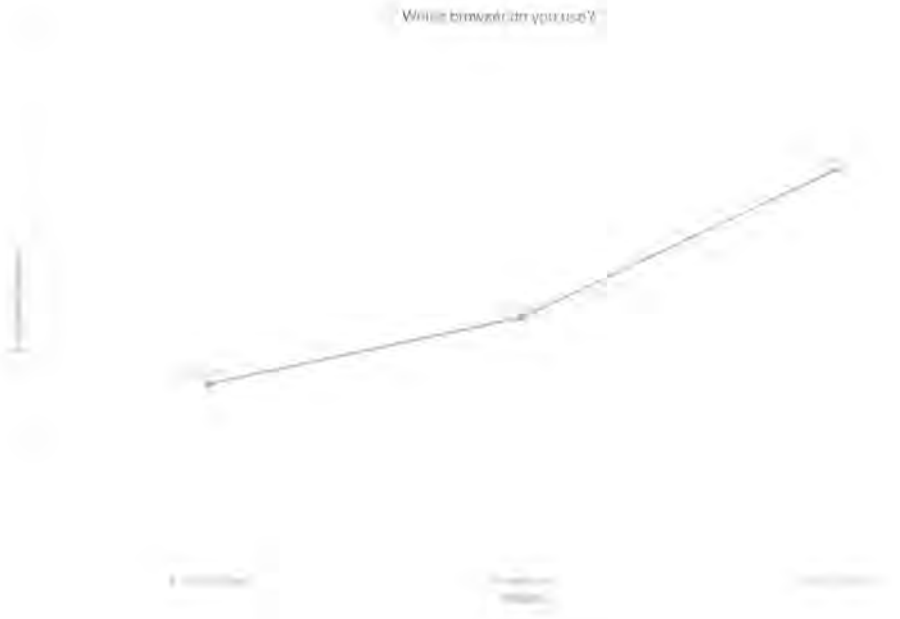


FIGURE 7.10: Internet browser choices

Brown noted that there are three main strategies to measure the reliability tests in which test-retest reliability, equivalent forms reliability and internal consistency reliability are distinguished [237]. Internal consistency reliability is the easiest among above mentioned three strategies because there is no need to administer two times or have the test into two forms. Internal consistency reliability has further three types through which test can be conducted. Kuder-Richardson formulas K-R20 and K-R21, split-half reliability and Cronbach's alpha are included in the internal consistency reliability.

User based evaluation was performed. The reliability of the questionnaire was checked based on four variables, namely Learnability, Ease of Use, Efficiency and Satisfaction. Table 7.1 exhibits the reliability of all the variables in the questionnaire.

The Cronbach's alpha values ($\alpha = 0.92$) which suggests that all the variables in the questionnaire were highly reliable. However, before accepting the overall reliability of

ıe questionnaire it was essential to check the alpha scores α for individual variables
; shown in table 7.1.

TABLE 7.1: Reliability statistics

Cronbach's Alpha	N of Items
0.92	4

The table 7.2 represents the reliability of each variable in the questionnaire.

TABLE 7.2: Item-total statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item Total Correlation	Cronbach's A if Item Dele
Learnability	9.7302	9.463	0.913	0.915
EaseOfUse	9.7302	9.463	0.912	0.914
Efficiency	9.873	9.269	0.919	0.916
Satisfaction	9.8571	9.136	0.92	0.926

The table 7.2 shows that all the variable individually suggests an item correlation from 0.912 to 0.920, which indicates that all the variables in the questionnaire meet the assumptions about the reliability in this study.

Moreover, in order to check the significance of the variables: Learnability, Ease of Use, Efficiency and Satisfaction, 'One sample t-test' was conducted. The table 7.3 shows the mean and the standard deviation (Std. Deviation).

TABLE 7.3: One-sample statistics

	N	Mean	Std. Deviation	Std. Error Mean
Satisfaction	63	3.3333	1.00000	0.12599
EaseOfUse	63	3.2333	1.02000	0.12599
Efficiency	63	3.1905	1.03514	0.13042
Learnability	63	3.0063	1.05614	0.13306

The table 7.3 shows that participants in this study were highly satisfied as satisfaction highest mean comes out to be $\mu = 3.33$ and standard deviation $\sigma = 1.00$

out of a total score of 5 on the questionnaire scale. Also the ease of use parameter proves to be highly accepted as the mean value for it was calculated as $\mu = 3.23$ having the standard deviation $\sigma = 1.02$.The efficiency of the system also proven to be proficient as the mean of it emanates as $\mu = 3.1905$ and the standard deviation $\sigma = 1.03$. Last but not the least, learnability is at the bottom having the mean value $\mu = 3.00$ having a slightly higher value of the standard deviation $\sigma = 1.05$.

7.5 Performance Optimization

Optimization is the selection of a best value (constrained by some criteria) amongst a set of available alternative values for a given problem. In our system time and error are the two variables which have to be optimized for performance by selecting the most suitable video resolution from the dataset. Visual quality is related to the resolution of the video, however, going beyond or below certain limits does not give satisfactory performance when used in video retrieval system. In our system, the constraints and objective function are all linear. Therefore, we need a linear optimization model for our system. The dataset has been prepared at all the different resolutions so that it can be useful in varied experiments of video retrieval. In our framework the optimum performance is observed at resolution 320x240 as depicted in the figure 7.11.

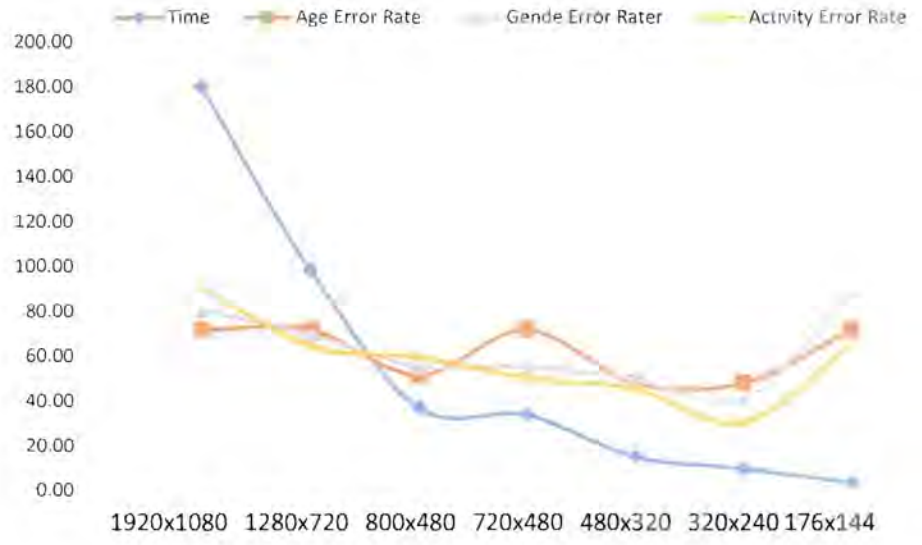


FIGURE 7.11: Performance optimization

In the figure 7.11 the error rate at resolution 320x240 and resolution 480x320 is almost same while there is improvement in time complexity for resolution 320x240. Hence, we have used the same resolution through out our experiments.

7.6 Summary

Graphical User Interface is a key element and entry point for user to interact with any system. It's designing and evaluation is essential for any computer systems. HCI plays an important role to remove the gap between the system and the users. In order to test and examine the proposed video retrieval approach, was developed to satisfy the user's information requirements. Usability testing approaches were incorporated for user based evaluation is design based and explains if the generated technique is efficient and meets the requirements of users. Time and error were two main variables used to optimise the performance of the designed system. Linear

optimisation model was used for the proposed system. As the system was optimized based on performance so, the optimized resolution was employed in the experiments.

Chapter 8

Conclusion and future work

The increase in production of digital images and video collections within the past decade has been overwhelming and will continue to show the same trend. It is therefore, imperative to devise techniques and methods to efficiently and effectively process and analyse the underlying semantics of the video content which can be used for video retrieval.

The researchers have focussed on the concepts of video structure analysis, feature extraction, content annotation, video tagging, indexing, querying and retrieval for the application in personal videos, news broadcast, sports and movies etc. However, some of these techniques are developed keeping focus on a single domain and over emphasize the issues of quality, hence requiring enormous processing and storage capabilities. Also, time constraints restrict the use of over complicated and/or sophisticated algorithms which can give plausible results in a controlled environment but do not fulfil the requirements of real time processing of real world data.

The research aimed at designing, implementing and evaluating the feature extraction, video indexing and retrieval approaches focussing on the content based feature extraction using low-level feature and high level semantics in order to minimise the errors and maximize the accuracy of the indexing system without requiring human

intervention which can significantly fill the gap identified through literature review while satisfying the imposed constraints of processing, storage, quality and retrieval times. The output is a methodology and proof of concept paving the way for extending similar research in the future. The potential application areas include but are not limited to the movie-industry, personal photo and video collections or internet archives.

Relevant literature review has identified a multitude of approaches used by other researchers to address the problem. Significant importance has been attributed to segmentation of video structure into shots and scenes; and extraction of features including static features like colour, texture and shapes and motion features. Annotation of videos based on content, context and by using integration of both has been implemented in conjunction with classification and all of them separately as well. The concept of indexed videos for efficient and effective querying and retrieval has been performed using segment based, object based and event based indexes.

For bridging the gap identified in the literature review, the video indexing framework has been proposed. The novelty of the approach is its coverage of the various components for feature extraction, video indexing and retrieval, which previously has been treated individually. The components are covered in the most feasible sequence they appear in video indexing and retrieval systems. We also covered the retrieval phase particularly from the user's perspective.

In order to ascertain the effectiveness and efficiency of computer algorithms, it is a standard practice to evaluate them on publically available standardized datasets. This allows a comprehensive comparison of the precision and recall of such algorithms and established state of the art and helps further the research. However to the best of our knowledge, no such dataset exist for the evaluation of video retrieval systems making it difficult to compare and evaluate different approaches covering various aspects of video indexing domains. To circumvent this limitation of performance evaluation of the proposed holistic framework, a comprehensive dataset

covering various aspects needed to support the framework at every intermediate stage has been produced by following the standard and baselines for video corpus generation proposed by the researchers in same context. The justification of creating the new dataset came from limitations posed by the existing datasets. Then we requested human experts to provide ground truth judgements for the videos to be used in our experiments. The manual annotation process got completed in two months time for the complete dataset. Cross validation was randomly performed on the manual annotation.

The proposed framework is designed to segment the video structure into smaller chunks which are independent of one another based on changing shot. Each chunk, which we call a segment, represents a transition from the previous segment into the next in the same video. Each of these segments is considered analogous to a word in text and cannot be used alone to elucidate the context or provide the semantic information. Transitions such as cut, fade in fade out, dissolve and wipe are available and implementable methods but based on our empirical analyses, the "cut" has been found to be the most significant transition. The "cut transition" is detected most efficiently using thresh-holding based approach which has little computational overhead and can be implemented in real time systems. Significant changes in the local neighbourhood histogram are observed when there is an abrupt change of the transition. This approach does not require adding/inserting additional frames to delineate different shots. However, fade in and fade out based shot boundary cannot be detected using this technique and requires further investigation to develop a feasible solution.

The analysis of the images begins with the fundamental process of feature extraction. All static images contain features such as colour, texture and geometric shapes which can be used to represent the semantics. Similar to the weights in text retrieval, the image features are represented as a feature vector whose dimensionality can be reduced if required. These features can be extracted in a global or local

fashion based on the requirements of the processing engine. In our experiments, local features have shown to be better than the global ones for content based image classification. The tradeoff between computational cost and efficiency has to be optimized to achieve the best combination and performance. The use of wavelets for texture analysis and image segmentation is a well established approach and using Haar wavelets for face detection is a robust and efficient method. The static frames from the video are extracted based on the frame rate, i.e., a 25 frames per second video segment will render 25 static frames. Each of these frames is subjected to Haar cascade based classification to locate faces which are later normalized to identify gender and age on the basis of geometric features. The performance of Haar in terms of time and accuracy is found to be significantly better than combination of canny edge detection and mean face template. We have assumed that the most significant feature for detecting human within a frame is their face. Experiment results show that this approach of Haar wavelet gives average precision and recall of 0.91 and 0.92 respectively which is comparatively reliable from the other tested approaches. The human faces which have been extracted the frames are then analysed for age detection using geometric features such as distance of eyes, nose and mouth etc. from reference points. This is useful to classify the subjects into three groups: baby, young and old. The average precision and the recall for identifying the young are 0.91 and 0.81 respectively followed by 0.85 and 0.84 for the baby and 0.82 and 0.78 for the old. We have used the modified algorithm for detecting geometric features to recognize gender of the subject. Features such as location of eye brows, width of face, size of eyes, location of the chin and prominence of Adams apple are landmarks used. Recognition of males have better precision of 0.89 vs. 0.70 for female while recall in female is 0.92 vs. 0.79 in male.

The activity is detected by binarization of the frame to separate foreground from background. As background is uniform and has lighter shade because our dataset comprises of indoor scenario, the foreground is primary subject which is human.

After removing the noise the image is converted into stick figure and lines and their angles are detected using Hough transform to establish the relationship between body segments. This intermediate result is use to classify the activities such as walk, wave, box and run by using Hidden Markov Model. Related activities such as walk and run show the overlapping results that are insignificant. The overall system performance has been optimized for minimum time required to extract and populate the features and give minimum errors in the detection against different resolution of the video. The trade of time vs. error has been significantly improved on the resolution range 320x240.

To assess the overall performance of the system the so called user based evaluation have also been performed. The graphical use interface has been designed as the front end and integrated into the backend which is video retrieval system. The user is given options of simple or advanced search to detect features of interest within the video base and retrieve annotated results. The browsing is an additional feature integrated into the GUI. The HCI based criteria i.e., user friendliness, effectiveness, learnability and usability for GUI have been evaluated which give plausible results.

8.1 Future Work

The conclusions drawn form the research guide to the open areas for the research which could be analysed in future. The essence of the finding for the future work are as follows: Shot boundary detection techniques could further be enhanced to fade in, fade out, wipe and dissolve. The algorithm for real time shot boundary detection could be optimised for high resolution and real time application. The identification of faces could be enhanced to detect partially visible or occluded faces. Age detection could further be modified to estimate age in decades by using machine learning techniques. Activity detection could be enhanced to detect activity against busy

background and the palette of activities could be expanded to support sporting events. Use of natural language queries for search system could be incorporated to give it a flexibility for search options. Embedding the annotation data within the movie meta data. The annotations can be used for security and censorship applications so that offensive media content can be automatically blocked by the media player.

Feature based dynamic intra video indexing encompasses different inter- connected constituents, which results in change of some components in overall system based on improvement of any specific component. Based on this, it is important to consider that the framework for indexing is designed to support improvements and on-going addition in it. The conclusions drawn from the research guide to the open areas for the research which could be analysed in future. The essence of the finding for the future work are as follows:

Shot boundary detection is the main entry point of the framework. Shot boundary detection could be of different types such as cut, fade in, fade out, wipe and dissolve. The framework incorporates cut detection as the main boundary detection feature. This can further be enhanced by including other categories of the shot change. In fade type of shot transition, one shot either fade in to another shot or fade out to another shot. A shot is classified as fade out if the existing frame is transformed into a black frame where as in fade in type of transition the main frame will appear from a black frame. Dissolve can be classified as fade-in and fade-out occurring synchronously. In wipe type of shot transition, one shot will be wiped using a virtual line in any temporal direction. Any of these gradual transition can be integrated in the framework. Camera zoom accompanied by light intensity leads to a non-translational occlusion and effect the detection process along with selection of inappropriate block for shot boundary detection. While histogram difference giving a promising result but higher-order colour models could be incorporated along with

the selection of hierarchal block to improve the performance. Use of The effectiveness of shot boundary detection yields in better performance of overall framework. The algorithm for real time shot boundary detection could be optimised for high resolution and real time application using graphics processing unit (GPU) which has become an integral part for high end computational processing. GPUs are normally considered to be powerful graphic engine, but they also got highly parallel and programmable architecture featuring topmost arithmetic and memory bandwidth, which makes them persuasive alternative to CPUs. They also prioritize throughput over latency which results into real time detection possible even at higher resolution aspect ratios. Embedding real time detection will lead to different scope of videos like live streams and these will be annotated at real time as well to be used in current indexing system.

The identification of faces could be enhanced to detect partially visible or occluded faces. It is suggested that a fine tuning from a coarse cycle will result into more effective detection. Surface representations and image pre-processing techniques are particularly well suitable to precise facial parts such as colour stabilisation for the mouth, edge detectors round the two eyes, slope depictions for the nose. Also by using compound subspaces focussed on particular areas of the face, robustness can be increased against partial occlusion, such as beards, glasses, caps or moustache.

The perfect age classifier would have the ability to point out accurate age which is completely human coordinated. Although, the unpremeditated age definer should have the ability to define different pose, expression, illumination, as well as image quality. Along with that, by minor changes if needed, the descriptor need to have the ability to be applied on numerous inhabitants. Small descriptors have the preference along with the sparse structures, if conceivable. Evidently the speedy calculation period is also another apprehension. It is considered that this work could be enhanced in future to tackle this challenge, in this research, more metric learning technique and features were used, it proceeded in such technique by shutting the

huge performance gulf majority. In future, 3D model-based arrangement could also be used with age detection, along with the feed forward models capability; it will efficiently be absorbed by plenty of specimens to end the restrictions and weaknesses of the existing techniques. For the representation of the marked improvement in age recognition the capable coupling has been attested which has become important in different visual domains.

For the human movements or actions understanding, this work shows the noticeable direction taken by the researcher. Such extremely pertinent subject is jut at its start, a great deal of work could be extended from it. Significant paths have also been defined by the researchers of this project for easy and complex activity identification. For feature extraction the main method has been selected as it is computationally economical because it is not iterative. For scale sensitivity control an unambiguous mechanism is also provided. It is depended on no a priori human prototype. For identification HMM has been used to show the specific action in the certain circumstances. The following work has the ability to extend in plenty of directions because the global image illustration had significantly given good outcomes. Along with that its extraction can be very inexpensive. But its application is limited to those situations, in which region of interest ROIs have resolute reliability. Furthermore, obstructions cannot be handled. The specific problems are handled by native representatives. Bag of feature was used in preliminary work, which can further be enhanced to use 3D and temporal association among patches. This work can also be enhance to use the queries of how to tackle with more larger image blocks. Vision based human action recognition have been used in this research which can be further enhanced to multi-modal approach for better recognition in some domains. Along with that, the settings like camera movement, background, person identity, and communication can also be considered in the future for better recognition. By giving recent scenario of art as well as inspiration of the comprehensive variety of applications, it has the ability to assist the strong human action recognition. It is

believed that in up-coming phase such tasks will be tackled. To get strong programmed recognition and elucidation of human action will be an enduring step of achievement to be fulfilled.

A system of video retrieval has been proposed; retrieval and indexing keywords are part of it. Although a perfect video retrieval system has not yet been established; which has the ability to retrieve statements like I want to watch a video shot, where President of United Sates Obama is shaking hand with prime minister of Pakistan Nawaz Sharif, it is also called complete textual descriptions. A connection among keywords is needed in these kinds of queries. Depending upon the portrayal of natural language, video scene classification is required. Such language is combined with the complementarity information taken out from each class of video scenes for classification assignment release. By searching the outcome of combined terminology among the documents, the errors of the classification can also be minimized. Classifications of visual scenes needs to be completely depended upon the descriptions of the machine generated outcomes. It is the impartial size of the machine generated descriptions the human remarks can also be incorporated as a counterpart, and vocabulary expressions can also be controlled in the future.

This work can also be extended for embedding the annotation data within the movie meta-data. Metadata attract the attention of many researcher for annotation purposes because of many reasons. Metadata and contents have simple and quick acceptance, encapsulation as well as transformation. By the help of compound metadata upper level granularity and upper level of detail is achieved, for example metadata narrative content making. In its capability for semi-automatic resource distribution, the open and regular interfaces are used as a standard. The tightly coupled metadata annotation assurances extraordinary performance. For the representation of content, accessible temporal and spatial metadata resolution has been used. The unambiguous connection between certain features as well as its function's role. For

the good user experience, the proper connections to the movie records are accumulated. For instance, there is a huge content of videos and by taking such steps the age suitable or irrelevant content can be restricted for the children so they can watch cartoons or other videos suitable for them. The main focus has been to alter the present records in the more controlled data groups. This transformation is to be made automated. Currently voluntarily, people assistance is needed to assist the classifiers for the movies and also for the recognition of the names and roles which are registered in the records of the library which needs to be fully automated in the future.

The annotations can be used for security and censorship applications so that offensive media content can be automatically blocked by the media player in the future. To handle the wrong circulation of copyright digital material, the Digital Rights Management (DRM) was presented. But it exasperated the users by the fact where DRM was irrelatively plunged by music industry because of the imposition of the strict restriction on the digital asset usage. In recent times the industry of entertainment has moved its responsiveness in restraining content shield in the limits of the domain of interoperable, in reference to recognize the growing users frustration. Fingerprinting content also known as perceptual or hashing robust, integrally has the divergence to the hash functions conventional cryptographic. Although the binary representations has a difference in the multimedia items ought to hash the same value, but as a perception they are considered same. On the consumer created way, that technology had quickly adopted admiration that is why it has been used daily for binding of the metadata as well as sifting out uploaded patent content. Before playing any video for the data annotations check, this methodology can be used to extend this work. To jam the offensive media content automatically by media player the annotations would be used for the security as well as censorship applications.

References

[1] Rudolfs Drillis, Renato Contini, and Maurice Bluestein. Body segment parameters. *Artificial limbs*, 8(1):44–66, 1964.

[2] Cisco2013. Cisco Visual Networking Index: Forecast and Methodology, 20122017.

[3] MS Lew, N Sebe, C Djeraba, and R Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia ...*, 2(1):1–19, 2006.

[4] Yuxin Peng and Chong-Wah Ngo. Hot event detection and summarization by graph modeling and matching. In *Image and Video Retrieval*, pages 257–266. Springer, 2005.

[5] Paul Over, George M Awad, Jon Fiscus, Martial Michel, Alan F Smeaton, and Wessel Kraaij. Trecvid 2009-goals, tasks, data, evaluation mechanisms and metrics. 2010.

[6] Alan F. Smeaton, Paul Over, and Aiden R. Doherty. Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, 114(4):411–418, April 2010.

[7] Alan F Smeaton, Paul Over, and Wessel Kraaij. High-level feature detection from video in trecvid: a 5-year retrospective of achievements. In *Multimedia content analysis*, pages 1–24. Springer, 2009.

[8] F. Pereira, a. Vetro, and T. Sikora. Multimedia Retrieval and Delivery: Essential Metadata Challenges and Standards. *Proceedings of the IEEE*, 96(4):721–744, April 2008.

[9] R Venkatesh Babu and KR Ramakrishnan. Compressed domain video retrieval using object and global motion descriptors. *Multimedia Tools and Applications*, 32(1):93–113, 2007.

[10] Alan F Smeaton. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Information Systems*, 32(4):545–559, 2007.

[11] Yuk Ying Chung, Waikwok Jess Chin, Xiaoming Chen, David Yu Shi, Eric Choi, and Fang Chen. Content-based video retrieval system using wavelet transform. *WSEAS Transactions on Circuits and Systems*, 6(2):259–265, 2007.

[12] VS Subrahmanian. *Principles of multimedia database systems*. Morgan Kaufmann Publishers Inc., 1998.

[13] Abdelsalam A Helal, Anupam Joshi, and Magdy Ahmed. *Video database systems: issues, products, and applications*, volume 5. Springer, 1997.

[14] JungHwan Oh and Kien A Hua. Efficient and cost-effective techniques for browsing and indexing large video databases. In *ACM SIGMOD Record*, volume 29, pages 415–426. ACM, 2000.

[15] Chuohao Yeo, Yong-wei Zhu, Qibin Sun, and Shih-fu Chang. A Framework for Sub-Window Shot Detection. *11th International Multimedia Modelling Conference*, pages 84–91, 2005.

[16] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. A Formal Study of Shot Boundary Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(2):168–186, February 2007.

[17] Siripinyo Chantamunee and Yoshihiko Gotoh. University of sheffield at trecvid 2007: Shot boundary detection and rushes summarisation. In *TRECVID*. Citeseer, 2007.

[18] Steven CH Hoi, Lawson LS Wong, and Albert Lyu. Chinese university of hongkong at trecvid 2006: Shot boundary detection and video search. In *TRECVid 2006 Workshop*, pages 76–86, 2006.

[19] Zhi-Cheng Zhao and An-Ni Cai. Shot boundary detection algorithm in compressed domain based on adaboost and fuzzy theory. In *Advances in Natural Computation*, pages 617–626. Springer, 2006.

[20] Sarah Victoria Porter. *Video segmentation and indexing using motion estimation*. PhD thesis, University of Bristol, 2004.

[21] Yuchou Chang, D. J. Lee, Yi Hong, and James Archibald. Unsupervised Video Shot Detection Using Clustering Ensemble with a Color Global Scale-Invariant Feature Transform Descriptor. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[22] X. Wu, Pong C. Yuen, C. Liu, and J. Huang. Shot Boundary Detection: An Information Saliency Approach. *2008 Congress on Image and Signal Processing*, pages 808–812, 2008.

[23] Xinbo Gao, Jie Li, and Yang Shi. A video shot boundary detection algorithm based on feature tracking. In *Rough Sets and Knowledge Technology*, pages 651–658. Springer, 2006.

[24] G Camara-Chavez, F Precioso, M Cord, S Phillip-Foliguet, and A de A Araujo. Shot boundary detection by a hierarchical supervised approach. In *Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on*, pages 197–200. IEEE, 2007.

[25] Hong Lu, Yap-Peng Tan, Xiangyang Xue, and Lide Wu. Shot boundary detection using unsupervised clustering and hypothesis testing. In *Communications, Circuits and Systems, 2004. ICCCAS 2004. 2004 International Conference on.* volume 2, pages 932–936. IEEE, 2004.

[26] Matthew Cooper, Ting Liu, and Eleanor Rieffel. Video segmentation via temporal pattern classification. *Multimedia, IEEE Transactions on*, 9(3):610–618, 2007.

[27] Liang Bai, Song-Yang Lao, Hai-Tao Liu, and Jiang Bu. Video shot boundary detection using petri-net. In *Machine Learning and Cybernetics, 2008 International Conference on*, volume 5, pages 3047–3051. IEEE, 2008.

[28] Chunxi Liu, Huiying Liu, Shuqiang Jiang, Qingming Huang, Yijia Zheng, and Weigang Zhang. Jdl at trecvid 2006 shot boundary detection. In *TRECVID 2006 Workshop*, 2006.

[29] Dingyuan Xia, Xuefei Deng, and Qingning Zeng. Shot Boundary Detection Based on Difference Sequences of Mutual Information. *Fourth International Conference on Image and Graphics (ICIG 2007)*, pages 389–394, August 2007.

[30] Kyong-Cheol Ko, Young Min Cheon, Gye-Young Kim, Hyung-Il Choi, Seong-Yoon Shin, and Yang-Won Rhee. Video shot boundary detection algorithm. In *Computer Vision, Graphics and Image Processing*, pages 388–396. Springer, 2006.

[31] Zuzana Cernekova, Ioannis Pitas, and Christophoros Nikou. Information theory-based shot cut/fade detection and video summarization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(1):82–91, 2006.

[32] a. Hanjalic. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105, 2002.

[33] Georges M Quŝenot, Daniel Moraru, and Laurent Besacier. Clips at trecvid: Shot boundary detection and feature detection. 2003.

[34] Cüneyt M Taskiran, Zygmunt Pizlo, Arnon Amir, Dulce Ponceleon, and Edward J Delp. Automated video program summarization using speech transcripts. *Multimedia, IEEE Transactions on*, 8(4):775–791, 2006.

[35] G Camara Chavez, F Precioso, M Cord, S Philipp-Foliguet, and Arnaldo de A Araujo. Shot boundary detection at trecvid 2006. *Proc. TREC Video Retrieval Eval*, page 1À8, 2006.

[36] Zhi-Cheng Zhao, Xing Zeng, Tao Liu, and An-Ni Cai. Bupt at trecvid 2007: Shot boundary detection. In *TRECVID*. Citeseer, 2007.

[37] Xue Ling, Li Chao, Li Huan, and Xiong Zhang. A General Method for Shot Boundary Detection. *2008 International Conference on Multimedia and Ubiquitous Engineering (mue 2008)*, pages 394–397, 2008.

[38] Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan F Smeaton. Trecvid 2005-an overview. 2005.

[39] Adam Herout, Vítezslav Beran, Michal Hradis, Igor Potucek, Pavel Zemcík, and Petr Chmelar. Trecvid 2007 by the brno group. In *TRECVID*, 2007.

[40] John S Boreczky and Lynn D Wilcox. A hidden markov model framework for video segmentation using audio and image features. In *Acoustics, Speech and*

Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, volume 6, pages 3741–3744. IEEE, 1998.

[41] Chi-chun Lo and Shuenn-jyi Wang. Video segmentation using a histogram-based fuzzy c-means clustering algorithm. *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*, 3:920–923, 2001.

[42] Uros Damnjanovic, Ebroul Izquierdo, and Marcin Grzegorzek. Shot boundary detection using spectral clustering. In *15th European Signal Processing Conference*, page 1779, 2007.

[43] Sarah De Bruyne, Davy Van Deursen, Jan De Cock, Wesley De Neve, Peter Lambert, and Rik Van de Walle. A compressed-domain approach for shot boundary detection on H.264/AVC bit streams. *Signal Processing: Image Communication*, 23(7):473–489, August 2008.

[44] H. Koumaras, G. Gardikis, G. Xilouris, E. Pallis, and a. Kourtis. Shot boundary detection without threshold parameters. *Journal of Electronic Imaging*, 15(2):020503, 2006.

[45] Chong-Wah Ngo. A robust dissolve detector by support vector machine. *Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA '03*, 3:283, 2003.

[46] Hun-Woo Yoo, Han-Jin Ryoo, and Dong-Sik Jang. Gradual shot boundary detection using localized edge blocks. *Multimedia Tools and Applications*, 28(3):283–300, 2006.

[47] Hari Sundaram and Shih-Fu Chang. Video scene segmentation using video and audio features. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, pages 1145–1148. IEEE, 2000.

[48] Alan Hanjalic, Reginald L Lagendijk, and Jan Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):580–588, 1999.

[49] Liang-Hua Chen, Yu-Chun Lai, and Hong-Yuan Mark Liao. Movie scene segmentation using background information. *Pattern Recognition*, 41(3):1056–1065, 2008.

[50] Z. Rasheed and M. Shah. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, 7(6):1097–1105, December 2005.

[51] W. Tavanapong and J. Zhou. Shot Clustering Techniques for Story Browsing. *IEEE Transactions on Multimedia*, 6(4):517–527, August 2004.

[52] Z. Rasheed and M. Shah. Scene detection in Hollywood movies and TV shows. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, 2:II–343–8.

[53] Li Zhao, Wei Qi, Yi-Jin Wang, Shi-Qiang Yang, and HongJiang Zhang. Video shot grouping using best-first model merging. In *Proceedings of SPIE*, volume 4315, page 262, 2001.

[54] Naveen Goela, Kevin Wilson, Feng Niu, Ajay Divakaran, and Isao Otsuka. An svm framework for genre-independent scene change detection. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 532–535. IEEE, 2007.

[55] Yun Zhai and Mubarak Shah. Video scene segmentation using markov chain monte carlo. *Multimedia, IEEE Transactions on*, 8(4):686–697, 2006.

[56] Yap-Peng Tan and Hong Lu. Model-based clustering and analysis of video scenes. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–617. IEEE, 2002.

[57] Zhiwei Gu, Tao Mei, Xian-Sheng Hua, Xiuqing Wu, and Shipeng Li. EMS: Energy Minimization Based Video Scene Segmentation. *Multimedia and Expo, 2007 IEEE International Conference on*, pages 520–523, July 2007.

[58] Yasuo Ariki, Masahito Kumano, and Kiyoshi Tsukada. Highlight scene extraction in real time from baseball live video. *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval - MIR '03*, page 209, 2003.

[59] Yun Zhai, Alper Yilmaz, and Mubarak Shah. Story segmentation in news videos using visual and text cues. In *Image and Video Retrieval*, pages 92–102. Springer, 2005.

[60] WH-M Hsu and Shih-Fu Chang. Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 2, pages 1091–1094. IEEE, 2004.

[61] Arnon Amir, Marco Berg, Shih-Fu Chang, Winston Hsu, Giridharan Iyengar, Ching-Yung Lin, Milind Naphade, Apostol Natsev, Chalapathy Neti, Harriet Nock, et al. Ibm research trecvid-2003 video retrieval system. *NIST TRECVID-2003*, 2003.

[62] Rong Yan and Alexander G Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4-5):445–484, 2007.

[63] John Adcock, Andreas Girgensohn, Matthew Cooper, Ting Liu, Lynn Wilcox, and Eleanor Rieffel. Fxpal experiments for trecvid 2004. *Proceedings of the TREC Video Retrieval Evaluation (TRECVID)*, pages 70–81, 2004.

[64] Alex Hauptmann, Robert V Baron, Ming-yu Chen, M Christel, Pinar Duygulu, C Huang, R Jin, W-H Lin, T Ng, and N Moraveji. Informedia at trecvid 2003: Analyzing and searching broadcast news video. Technical report, DTIC Document, 2004.

[65] Alex Hauptmann, MY Chen, Mike Christel, C Huang, Wei-Hao Lin, T Ng, Norman Papernick, A Velivelli, Jie Yang, Rong Yan, et al. Confounded expectations: Informedia at trecvid 2004. In *Proc. of TRECVID*, 2004.

[66] Colum Foley, Cathal Gurrin, Gareth JF Jones, Hyowon Lee, Sinéad McGivney, Noel E O'Connor, Sorin Sav, Alan F Smeaton, and Peter Wilkins. Trecvid 2005 experiments at dublin city university. Technical report, NIST, 2005.

[67] Rene Visser, Nicu Sebe, and Erwin Bakker. Object recognition for video retrieval. In *Image and Video Retrieval*, pages 262–270. Springer, 2002.

[68] Josef Sivic, Mark Everingham, and Andrew Zisserman. Person spotting: video shot retrieval for face sets. In *Image and Video Retrieval*, pages 226–236. Springer, 2005.

[69] Duy-Dinh Le, Shinichi Satoh, and Michael E Houle. Face retrieval in broadcasting news video by fusing temporal and intensity information. In *Image and Video Retrieval*, pages 391–400. Springer, 2006.

[70] Huiping Li and David Doermann. Video indexing and retrieval based on recognized text. In *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 245–248. IEEE, 2002.

[71] Ronan Fablet, Patrick Bouthemy, and Patrick Pérez. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *Image Processing, IEEE Transactions on*, 11(4):393–407, 2002.

[72] Yu-Fei Ma and Hong-Jiang Zhang. Motion texture: a new motion based video representation. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 548–551. IEEE, 2002.

[73] Till Quack, Vittorio Ferrari, and Luc Van Gool. Video mining with frequent itemset configurations. In *Image and Video Retrieval*, pages 360–369. Springer, 2006.

[74] W. Chen, H.J. Meng, and H. Sundaram. A fully automated content-based video search engine supporting spatiotemporal queries. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):602–615, 1998.

[75] Faisal I Bashir, Ashfaq A Khokhar, and Dan Schonfeld. Real-time motion trajectory-based indexing and retrieval of video sequences. *Multimedia, IEEE Transactions on*, 9(1):58–65, 2007.

[76] William Chen and Shih-Fu Chang. Motion trajectory matching of video objects. In *Electronic Imaging*, pages 544–553. International Society for Optics and Photonics, 1999.

[77] Young-kee Jung, Kyu-won Lee, and Yo-sung Ho. Content-based event retrieval using semantic scene interpretation for automated traffic surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 2(3):151–163, 2001.

[78] Chih-Wen Su, H-YM Liao, Hsiao-Rong Tyan, Chia-Wen Lin, Duan-Yu Chen, and Kuo-Chin Fan. Motion flow-based video retrieval. *Multimedia, IEEE Transactions on*, 9(6):1193–1201, 2007.

[79] Jun-wei Hsieh, Shang-li Yu, and Yung-sheng Chen. Motion-based video retrieval by trajectory matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(3):396–409, March 2006.

[80] a. Del Bimbo, E. Vicario, and D. Zingoni. Symbolic description and visual querying of image sequences using spatio-temporal logic. *IEEE Transactions on Knowledge and Data Engineering*, 7(4):609–622, 1995.

[81] Chikashi Yajimat Yoshihiro Nakanishi and Katsumi Tanaka. Querying video data by spatio-temporal relationships of moving object traces. In *Visual and Multimedia Information Management: IFIP TC 2/WG 2.6 Sixth Working Conference on Visual Database Systems, May 29-31, 2002, Brisbane, Australia*, page 357. Kluwer Academic Pub, 2002.

[82] Jinhui Tang, Xian-Sheng Hua, Meng Wang, Zhiwei Gu, Guo-Jun Qi, and Xiuqing Wu. Correlative linear neighborhood propagation for video annotation. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 39(2):409–16, April 2009.

[83] Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*, page 421, 2006.

[84] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 17–26. ACM, 2007.

[85] SL Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–1002. IEEE, 2004.

[86] Milind R Naphade and John R Smith. On the detection of semantic concepts at trecvid. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667. ACM, 2004.

[87] Yan Song, Guo-jun Qi, Xian-sheng Hua, Li-rong Dai, and Ren-hua Wang. Video Annotation by Active Learning and Semi-Supervised Ensembling. *2006 IEEE International Conference on Multimedia and Expo*, pages 933–936, July 2006.

[88] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and Jesse S Jin. A unified framework for semantic shot classification in sports video. *Multimedia, IEEE Transactions on*, 7(6):1066–1083, 2005.

[89] Xipeng Shen, Matthew Boutell, Jiebo Luo, and Christopher Brown. Multilabel machine learning and its application to semantic scene classification. In *Electronic Imaging 2004*, pages 188–199. International Society for Optics and Photonics, 2003.

[90] L. Hollink and M. Worring. Building a visual ontology for video retrieval. *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, page 479, 2005.

[91] Y. Wu, B.L. Tseng, and J.R. Smith. Ontology-based multi-classification learning for video concept detection. *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, pages 1003–1006, 2004.

[92] John R Smith, Milind Naphade, and Apostol Natsev. Multimedia semantic indexing using model vectors. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–445. IEEE, 2003.

[86] Milind R Naphade and John R Smith. On the detection of semantic concepts at trecvid. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 660–667. ACM, 2004.

[87] Yan Song, Guo-jun Qi, Xian-sheng Hua, Li-rong Dai, and Ren-hua Wang. Video Annotation by Active Learning and Semi-Supervised Ensembling. *2006 IEEE International Conference on Multimedia and Expo*, pages 933–936. July 2006.

[88] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and Jesse S Jin. A unified framework for semantic shot classification in sports video. *Multimedia, IEEE Transactions on*, 7(6):1066–1083, 2005.

[89] Xipeng Shen, Matthew Boutell, Jiebo Luo, and Christopher Brown. Multilabel machine learning and its application to semantic scene classification. In *Electronic Imaging 2004*, pages 188–199. International Society for Optics and Photonics, 2003.

[90] L. Hollink and M. Worring. Building a visual ontology for video retrieval. *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, page 479, 2005.

[91] Y. Wu, B.L. Tseng, and J.R. Smith. Ontology-based multi-classification learning for video concept detection. *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, pages 1003–1006, 2004.

[92] John R Smith, Milind Naphade, and Apostol Natsev. Multimedia semantic indexing using model vectors. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–445. IEEE, 2003.

[93] Wei Jiang, Shih-Fu Chang, and Alexander C Loui. Active context-based concept fusionwith partial user labels. In *Image Processing, 2006 IEEE International Conference on*, pages 2917–2920. IEEE, 2006.

[94] Jianping Fan, Hangzai Luo, and Ahmed K Elmagarmid. Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 13(7):974–92, July 2004.

[95] Rong Yan and Milind Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 657–663. IEEE, 2005.

[96] Xun Yuan, Xian-Sheng Hua, Meng Wang, and Xiu-Qing Wu. Manifold-ranking based video concept detection on large database and feature pool. *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*, page 623, 2006.

[97] Meng Wang, Yan Song, Xun Yuan, Hong-Jiang Zhang, Xian-Sheng Hua, and Shipeng Li. Automatic video annotation by semi-supervised learning with kernel density estimation. *Proceedings of the 14th annual ACM international conference on Multimedia - MULTIMEDIA '06*, page 967, 2006.

[98] Meng Wang, Xian-sheng Hua, Jinhui Tang, and Richang Hong. Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE Transactions on Multimedia*, 11(3):465–476, April 2009.

[99] Ralph Ewerth and Bernd Freisleben. Semi-supervised learning for semantic video retrieval. *Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07*, pages 154–161, 2007.

[100] Yan Song, Xian-Sheng Hua, Li-Rong Dai, and Meng Wang. Semi-automatic video annotation based on active learning with multiple complementary predictors. *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval - MIR '05*, page 97, 2005.

[101] Yan Song, Xian-Sheng Hua, Guo-Jun Qi, Li-Rong Dai, Meng Wang, and Hong-Jiang Zhang. Efficient semantic annotation method for indexing large personal video database. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*, (4):289, 2006.

[102] Guojun Lu. *Multimedia database management systems.* Artech House Norwood, 1999.

[103] Roland Tusch, Harald Kosch, and Lázló Böszörményi. Videx: an integrated generic video indexing approach. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 448–451. ACM, 2000.

[104] Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.

[105] Gulrukh Ahanger and Thomas D. C. Little. Data semantics for improving retrieval performance of digital news video systems. *Knowledge and Data Engineering, IEEE Transactions on*, 13(3):352–360, 2001.

[106] Ahmet Ekin, A Murat Tekalp, and Rajiv Mehrotra. Integrated semantic-syntactic video modeling for search and browsing. *Multimedia, IEEE Transactions on*, 6(6):839–851, 2004.

[107] Riccardo Leonardi and Pierangelo Migliorati. Semantic indexing of multimedia documents. *MultiMedia, IEEE*, 9(2):44–51, 2002.

[108] HongJiang Zhang. Content-based video browsing and retrieval. *Handbook of Internet and multimedia systems and applications*, pages 255–280, 1999.

[109] Carl Espen Lauter. Video database systems. *21st Computer Science Seminar*, 2005.

[110] Dulce Ponceleon, Savitha Srinivasan, Arnon Amir, Dragutin Petkovic, and Dan Diklic. Key to effective video retrieval: effective cataloging and browsing. In *Proceedings of the sixth ACM international conference on Multimedia*, pages 99–107. ACM, 1998.

[111] Cees GM Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005.

[112] Daniel DeMenthon, Longin Jan Latecki, Azriel Rosenfeld, and Marc Vuilleumier Stückelberg. Relevance ranking of video data using hidden markov model distances and polygon simplification. In *Advances in Visual Information Systems*, pages 49–61. Springer, 2000.

[113] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655, 1998.

[114] Changick Kim and Jenq-Neng Hwang. Fast and automatic video object segmentation and tracking for content-based applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(2):122–129, 2002.

[115] Todd Schoepflin, Vikram Chalana, David R Haynor, and Yongmin Kim. Video object tracking with a sequential hierarchy of template deformations. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(11):1171–1182, 2001.

[116] Lihi Zelnik-Manor and Michal Irani. Event-based analysis of video. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–123. IEEE, 2001.

[117] Noboru Babaguchi, Yoshihiko Kawai, and Tadahiro Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *Multimedia, IEEE Transactions on*, 4(1):68–75, 2002.

[118] Masayoshi Teraguchi, Ken Masumitsu, Tomio Echigo, SI Sekiguchi, and Minoru Etoh. Rapid generation of event-based indexes for personalized video digests. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 1041–1044. IEEE, 2002.

[119] Chuan Wu, Yu-Fei Ma, Hong-Jiang Zhan, and Yu-Zhuo Zhong. Events recognition by semantic inference for sports video. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 805–808. IEEE, 2002.

[120] Baoxin Li and M Ibrahim Sezan. Event detection and summarization in sports video. In *Content-Based Access of Image and Video Libraries, 2001.(CBAIVL 2001). IEEE Workshop on*, pages 132–138. IEEE, 2001.

[121] Anil Kokaram and Perrine Delacourt. A new global motion estimation algorithm and its application to retrieval in sports events. In *Multimedia Signal Processing, 2001 IEEE Fourth Workshop on*, pages 251–256. IEEE, 2001.

[122] José M Martínez, Rob Koenen, and Fernando Pereira. Mpeg-7: the generic multimedia content description standard, part 1. *Multimedia, IEEE*, 9(2):78–87, 2002.

[123] Philippe Salembier, Joan Llach, and Luis Garrido. Visual segment tree creation for mpeg-7 description schemes. *Pattern Recognition*, 35(3):563–579, 2002.

[124] Mohamed Abdel-Mottaleb and Santhana Krishnamachari. Multimedia descriptions based on mpeg-7: extraction and applications. *Multimedia, IEEE Transactions on*, 6(3):459–468, 2004.

[125] Frank Nack and Adam T Lindsay. Everything you wanted to know about mpeg-7. 1. *Multimedia, IEEE*, 6(3):65–77, 1999.

[126] F Pereira. Mpeg-7 requirements document v. 14. international, international organisation for standardisation, coding of moving pictures and audio. *International Organisation For Standardisation, Coding of Moving Pictures and Audio ISO/IEC JTC*, 1, 2001.

[127] Shih-Fu Chang, Thomas Sikora, and A Purl. Overview of the mpeg-7 standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):688–695, 2001.

[128] Arun Hampapur and Ramesh Jain. Video data management systems: Metadata and architecture., 1998.

[129] Weiming Hu, Dan Xie, Zhouyu Fu, Wenrong Zeng, and Steve Maybank. Semantic-based surveillance video retrieval. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 16(4):1168–81, April 2007.

[130] Josef Sivic and Andrew Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, pages 127–144. Springer, 2006.

[131] Yusuf Aytar, Mubarak Shah, and Jiebo Luo. Utilizing semantic word similarity measures for video retrieval. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[132] Lyndon S Kennedy, Apostol Paul Natsev, and Shih-Fu Chang. Automatic discovery of query-class-dependent models for multimodal search. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 882–891. ACM, 2005.

[133] Rong Yan, Jun Yang, and Alexander G. Hauptmann. Learning query-class dependent weights in automatic video retrieval. *Proceedings of the 12th annual ACM international conference on Multimedia - MULTIMEDIA '04*, page 548, 2004.

[134] Rong Yan and Alexander G. Hauptmann. Probabilistic latent query analysis for combining multiple retrieval sources. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, page 324, 2006.

[135] Rainer Lienhart. A system for effortless content annotation to unfold the semantics in videos. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 45–49. IEEE, 2000.

[136] Yi Wu, Yueting Zhuang, and Yunhe Pan. Content-based video similarity model. *Proceedings of the eighth ACM international conference on Multimedia - MULTIMEDIA '00*, pages 465–467, 2000.

[137] a. Anjulan and N. Canagarajah. A Unified Framework for Object Retrieval and Mining. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1):63–76, January 2009.

[138] Wen-Nung Lie and Wei-Chuan Hsiao. Content-based video retrieval based on object motion trajectory. In *Multimedia Signal Processing, 2002 IEEE Workshop on*, pages 237–240. IEEE, 2002.

[139] C.G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring. Adding Semantics to Detectors for Video Retrieval. *IEEE Transactions on Multimedia*, 9(5):975–986, August 2007.

[140] Shi-Yong Neo, Jin Zhao, Min-Yen Kan, and Tat-Seng Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Image and Video Retrieval*, pages 143–152. Springer, 2006.

[141] Bongnam Kang Hai Wang and Daijin Kim. Pfw: a face database in the wild for studying face identification and verification in uncontrolled environment. In *ACPR*, 2013.

[142] Luo Jie, Barbara Caputo, and Vittorio Ferrari. Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation. 2009.

[143] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 178–178. IEEE, 2004.

[144] Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[145] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001.

[146] Henry A Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(1):23–38, 1998.

[147] Ferdinando S Samaria and Andy C Harter. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142. IEEE, 1994.

[148] ChaLearn. Chalearn looking at people.

[149] Li Liu and Ling Shao. Learning discriminative representations from rgb-d video data. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2013.

[150] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 53–60, 2013.

[151] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative video pattern search for efficient action detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(9):1728–1743, 2011.

[152] Shuai Zheng, Junge Zhang, Kaiqi Huang, Ran He, and Tieniu Tan. Robust view transformation model for gait recognition. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2073–2076. IEEE, 2011.

[153] S. Singh, S.A. Velastin, and H. Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In

Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, pages 48–55, 2010.

[154] Ivan Laptev and Patrick Pérez. Retrieving actions in movies. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[155] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *The Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pages 1395–1402, 2005.

[156] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[157] Di Zhong, HongJiang Zhang, and Shih-Fu Chang. Clustering methods for video browsing and annotation. In *Electronic Imaging: Science & Technology*, pages 239–246. International Society for Optics and Photonics, 1996.

[158] Wei Jyh Heng and King Ngi Ngan. Shot boundary refinement for long transition in digital video sequence. *Multimedia, IEEE Transactions on*, 4(4):434–445, 2002.

[159] Alan Hanjalic. Shot-boundary detection: unraveled and resolved? *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(2):90–105, 2002.

[160] Jun Yu and Mandyam D Srinath. An efficient method for scene cut detection. *Pattern Recognition Letters*, 22(13):1379–1391, 2001.

[161] TRECVID 2012. Guidelines for TRECVID 2012.

[162] TRECVID 2013. Guidelines for TRECVID 2013.

[163] Silvia Pfeiffer. Pause concepts for audio segmentation at different semantic levels. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 187–193. ACM, 2001.

[164] Mrinal K Mandal, F Idris, and Sethuraman Panchanathan. A critical evaluation of image and video indexing techniques in the compressed domain. *Image and Vision Computing*, 17(7):513–529, 1999.

[165] Mee-Sook Lee, Yun-Mo Yang, and Seong-Whan Lee. Automatic video parsing using shot boundary detection and camera operation analysis. *Pattern Recognition*, 34(3):711–719, 2001.

[166] Nevenka Dimitrova, Yong Rui, and Ishwar K Sethi. Media content management. *Interactive Multimedia Systems*, page 1, 2002.

[167] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[168] Herbert A Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[169] Brian D Ripley. *Modern applied statistics with S*. Springer, 2002.

[170] Michael J Swain and Dana H Ballard. Indexing via color histograms. In *Active Perception and Robot Vision*, pages 261–273. Springer, 1992.

[171] Wenjing Jia, Huaifeng Zhang, Xiangjian He, and Qiang Wu. Gaussian weighted histogram intersection for license plate classification. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 574–577. IEEE.

[172] Ka-Man Wong, Chun-Ho Cheung, and Lai-Man Po. Merged-color histogram for color image retrieval. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages 949–952. IEEE, 2002.

[173] Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(5):522–529, 1995.

[174] S.K. Nayar and R.M. Bolle. Reflectance ratio: A photometric invariant for object recognition. In *Computer Vision, 1993. Proceedings., Fourth International Conference on*, pages 280–285, 1993.

[175] Theo Gevers and WM Smeulders. Color constant ratio gradients for image segmentation and similarity of texture objects. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–18. IEEE, 2001.

[176] Wenjing Jia, Huaifeng Zhang, Xiangjian He, and Qiang Wu. Image matching using colour edge cooccurrence histograms. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 3, pages 2413–2419. IEEE, 2006.

[177] Sir Ronald Aylmer Fisher, Statistiker Genetiker, Ronald Aylmer Fisher, Statistician Genetician, Great Britain, Ronald Aylmer Fisher, and Statisticien Généticien. *Statistical methods for research workers*, volume 14. Oliver and Boyd Edinburgh, 1970.

[178] Sameer K Antani, Mukil Natarajan, Jonathan L Long, L Rodney Long, and George R Thoma. Developing a comprehensive system for content-based retrieval of image and text data from a national survey. In *Medical Imaging*, pages 152–161. International Society for Optics and Photonics, 2005.

[179] Yong Rui, Alfred C She, and Thomas S Huang. Modified fourier descriptors for shape representation-a practical approach. In *Proceedings First Int'l Workshop Image Databases and Multi Media Search*, volume 22, page 23. Citeseer, 1996.

[180] Cha Zhang and Zhengyou Zhang. A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research, 2010.

[181] DN Chandrappa and M Ravishankar. Automatic face recognition in a crowded scene using multi layered clutter filtering and independent component analysis. In *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*, pages 552–556. IEEE, 2012.

[182] Chen Aiping, Pan Lian, Tong Yaobin, and Ning Ning. Face detection technology based on skin color segmentation and template matching. In *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*, volume 2, pages 708–711. IEEE, 2010.

[183] S Phimoltares, C Lursinsap, and K Chamnongthai. Face detection and facial feature localization without considering the appearance of image context. *Image and Vision Computing*, 25(5):741–753, 2007.

[184] Moi Hoon Yap, Hassan Ugail, and Reyer Zwiggelaar. Facial analysis for real-time application: A review in visual cues detection techniques. *Journal of Communication and Computer*, 9:1231–1241, 2012.

[185] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[186] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1955–1976, 2010.

[187] Narayanan Ramanathan, Rama Chellappa, and Soma Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009.

[188] Li Liu, Jianming Liu, and Jun Cheng. Age-group classification of facial images. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, pages 693–696. IEEE, 2012.

[189] Young H Kwon and Niels Da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999.

[190] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):442–455, 2002.

[191] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2234–2240, 2007.

[192] Jinli Suo, Tianfu Wu, Songchun Zhu, Shiguang Shan, Xilin Chen, and Wen Gao. Design sparse features for age estimation using hierarchical face model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[193] Narayanan Ramanathan and Rama Chellappa. Face verification across age progression. *Image Processing, IEEE Transactions on*, 15(11):3349–3361, 2006.

[194] Yih-Chuan Lin and Shen-Chuan Tai. A fast linde-buzo-gray algorithm in image vector quantization. *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, 45(3):432–435, 1998.

[195] Beatrice Alexandra Golomb, DT Lawrence, and TJ Sejnowski. Sexnet: A neural network identifies sex from human faces. *Advances in neural information processing systems*, 3:572–577, 1991.

[196] Shinichi Tamura, Hideo Kawai, and Hiroshi Mitsumoto. Male/female identification from 8× 6 very low resolution face images by neural network. *Pattern Recognition*, 29(2):331–335, 1996.

[197] Srinvis Gutta and Harry Wechsler. Gender and ethnic classification of human faces using hybrid classifiers. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 6, pages 4084–4089. IEEE, 1999.

[198] Hervé Abdi, Dominique Valentin, Betty Edelman, and Alice J O'Toole. More about the difference between men and women: evidence from linear neural network and the principal-component approach. *PERCEPTION-LONDON-*, 24:539–539, 1995.

[199] Abhishek Kumar, Karan Rawat, and Deepak Gupta. An advance approach of pca for gender recognition. In *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*, pages 59–63. IEEE, 2013.

[200] Baback Moghaddam and Ming-Hsuan Yang. Gender classification with support vector machines. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 306–311. IEEE, 2000.

[201] Brunelli Poggio, R Brunelli, and T Poggio. Hyberbf networks for gender classification. 1992.

[202] M Nazir, Muhammad Ishtiaq, Anab Batool, M Arfan Jaffar, and Anwar M Mirza. Feature selection for efficient gender classification. In *WSEAS International conference*, 2010.

[203] Ivan Laptev and Tony Lindeberg. Local descriptors for spatio-temporal recognition. In *Spatial Coherence for Visual Motion Analysis*, pages 91–103. Springer, 2006.

[204] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[205] James E Cutting and Lynn T Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, 9(5):353–356, 1977.

[206] Maylor K. Leung and Yee-Hong Yang. First sight: A human body outline labeling system. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(4):359–377, 1995.

[207] Hsuan-Sheng Chen, Hua-Tsung Chen, Yi-Wen Chen, and Suh-Yin Lee. Human action recognition using star skeleton. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 171–178. ACM, 2006.

[208] Hironobu Fujiyoshi and Alan J Lipton. Real-time human motion analysis by image skeletonization. In *Applications of Computer Vision, 1998. WACV'98. Proceedings., Fourth IEEE Workshop on*, pages 15–21. IEEE, 1998.

[209] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785, 1997.

[210] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[211] David Liu and Tsuhan Chen. A topic-motion model for unsupervised video object discovery. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[212] David Cunado, Mark S Nixon, and John N Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, 2003.

[213] Amit Kale, Aravind Sundaresan, AN Rajagopalan, Naresh P Cuntoor, Amit K Roy-Chowdhury, Volker Kruger, and Rama Chellappa. Identification of humans using gait. *Image Processing, IEEE Transactions on*, 13(9):1163–1173, 2004.

[214] Aravind Sundaresan, Amit RoyChowdhury, and Rama Chellappa. A hidden markov model based framework for recognition of humans from gait sequences. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–93. IEEE, 2003.

[215] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001.

[216] Ramprasad Polana and Randal Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, pages 77 82. IEEE, 1994.

[217] Ismail Haritaoglu, David Harwood, and Larry S Davis. Ghost: A human body part labeling system using silhouettes. In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 1, pages 77–82. IEEE, 1998.

[218] Richard C Staunton. An analysis of hexagonal thinning algorithms and skeletal shape representation. *Pattern Recognition*, 29(7):1131–1146, 1996.

[219] David A Winter. *Biomechanics and motor control of human movement*. Wiley. com, 2009.

[220] Jang-Hee Yoo, Doosung Hwang, and Mark S Nixon. Gender classification in human gait using support vector machine. In *Advanced concepts for intelligent vision systems*, pages 138–145. Springer, 2005.

[221] Linda Shapiro and George Stockman. Computer vision. 2001. *ed: Prentice Hall*, 2001.

[222] Richard O Duda and Peter E Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[223] Paul VC Hough. Machine analysis of bubble chamber pictures. In *International Conference on High Energy Accelerators and Instrumentation*, volume 73, 1959.

[224] I Holmes and GM Rubin. An expectation maximization algorithm for training hidden substitution models. *Journal of molecular biology*, 317(5):753–764, 2002.

[225] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260-269, 1967.

[226] Martha C Polson and J Jeffrey Richardson. *Foundations of intelligent tutoring systems.* Psychology Press, 2013.

[227] Raymonde Guindon. *Cognitive science and its applications for human-computer interaction.* Psychology Press, 2013.

[228] Alan Dix. *Human computer interaction.* Pearson Education, 2004.

[229] Michael G Christel and Ronald M Conescu. Mining novice user activity with trecvid interactive retrieval tasks. In *Image and Video Retrieval*, pages 21–30. Springer, 2006.

[230] Michael G Christel, Chang Huang, Neema Moraveji, and Norman Papernick. Exploiting multiple modalities for interactive video retrieval. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 3, pages iii–1032. IEEE, 2004.

[231] Marcel Worring, Cees GM Snoek, O De Rooij, GP Nguyen, and AWM Smeulders. The mediamill semantic video search engine. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1213. IEEE, 2007.

[232] C.G.M. Snoek, M. Worring, D.C. Koelma, and a.W.M. Smeulders. A Learned Lexicon-Driven Paradigm for Interactive Video Retrieval. *IEEE Transactions on Multimedia*, 9(2):280–292, February 2007.

[233] William W Agresti. The conventional software life-cycle model: its evolution and assumptions. *New Paradigms for Software Development*, pages 2–5, 1986.

[234] M.N. Asghar, F. Hussain, and R. Manton. A framework for feature based dynamic intravideo indexing. In *Automation and Computing (ICAC), 2013 19th International Conference on*, pages 1–6, 2013.

[235] Jakob Nielsen. Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414. ACM, 1994.

[236] Jeffrey Rubin and Dana Chisnell. *Handbook of usability testing: howto plan, design, and conduct effective tests.* Wiley. com, 2008.

[237] James Dean Brown. *Using surveys in language programs.* Cambridge University Press, 2001.

University of Bedfordshire

irac

PHOTOGRAPHY / VIDEO RECORDING CONSENT FORM – ADULT

Please read and ensure you are fully aware of this document before you sign, should you choose to.

To comply with the UK Data Protection Act, your permission is needed before we are able to use an image or any near likeness of you. No image or likeness will be used without your permission.
If it is given, you may remove your permission at any time by contacting the person named at the bottom of this document.
This form must not be signed by a minor. Permission for a minor must come from a legal guardian or parent.

Name.....

Location / activity .....
(University of Bedfordshire, Video Recording for Research Dataset)

Address
(home).....
.....

Postcode..... Contact telephone No.....

I give permission as the named person above for my likeness to be used for the purposes of University of Bedfordshire's internal and external research publications, newsletters, presentations and research articles.

Signed.....

Print.....

Date.....

PLEASE RETURN THIS FORM TO .....

Muhammad Nabeel Asghar

IRAC,
University of Bedfordshire.

University of Bedfordshire

irac

PHOTOGRAPHY / VIDEO RECORDING CONSENT FORM – ADULT

Please read and ensure you are fully aware of this document before you sign, should you choose to.

To comply with the UK Data Protection Act, your permission is needed before we are able to use an image or any near likeness of you. No image or likeness will be used without your permission.
If it is given, you may remove your permission at any time by contacting the person named at the bottom of this document.
This form must not be signed by a minor. Permission for a minor must come from a legal guardian or parent.

Name.....

Location / activity .....
(University of Bedfordshire, Video Recording for Research Dataset)

Address
(home).....
.....

Postcode..... Contact telephone No.....

I give permission as the named person above for my likeness to be used for the purposes of University of Bedfordshire's internal and external research publications, newsletters, presentations and research articles.

Signed.....

Print.....

Date.....

PLEASE RETURN THIS FORM TO .....

Muhammad Nabeel Asghar

IRAC,
University of Bedfordshire.

Dear participant

I am conducting this research solely for academic purposes. This survey is intended to assess the usability of video indexing system.The information you provide will be part of aggregated data and your annotations and propositions will not be individually identified. All feedback will be confidential and will only used for research purposes.

Please Tick the relevant ☑

What is your name (Optional) _____

What is your gender Male ☐ Female ☐

What is your age? 18-25 ☐ 26-39 ☐ 40-60 ☐ Above 60 ☐

What is your highest level of education? College ☐ Undergraduate ☐ Postgraduate ☐

How do you rate your computer skills? Beginer ☐ Intermediate ☐ Advance ☐

Which browser do you use Internet Explorer ☐ Mozilla Firefox ☐ Google Chrome ☐

(Write in box)
Other _____

Please Tick the relevant ☑

	Strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
I would like to use system again	☐	☐	☐	☐	☐
I find system to be easy to use	☐	☐	☐	☐	☐
I don't think I need special training to use system.	☐	☐	☐	☐	☐
I think most people will easily use system.	☐	☐	☐	☐	☐
I feel very confident while using the system	☐	☐	☐	☐	☐

Please Tick the relevant ☑

	strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
Position of video on screen is good?	☐	☐	☐	☐	☐
Consistent use of terms throughout the system	☐	☐	☐	☐	☐
Prompt for input are relevant?	☐	☐	☐	☐	☐
Designed for all level of users	☐	☐	☐	☐	☐
The information organized on the screen is optimised?	☐	☐	☐	☐	☐

Please Tick the relevant ☑

	strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
Overall I am satisfied with how easy the system is.	☐	☐	☐	☐	☐
Error messages were properly notified?	☐	☐	☐	☐	☐
It was simple to use system.	☐	☐	☐	☐	☐
I can effectively complete my task using the system.	☐	☐	☐	☐	☐
I was able to complete my tasks in reasonable time.	☐	☐	☐	☐	☐

Please Tick the relevant ☑

	strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
The system is visually appealing.	☐	☐	☐	☐	☐
The overall organization of the system is easily understandable.	☐	☐	☐	☐	☐
System is well designed.	☐	☐	☐	☐	☐
Terminology used in the system is clear.	☐	☐	☐	☐	☐
The result met my expectation.	☐	☐	☐	☐	☐
I would like to use this system again in future.	☐	☐	☐	☐	☐
Overall, the system is easy to use.	☐	☐	☐	☐	☐

Dear participant

I am conducting this research solely for academic purposes. This survey is intended to assess the usability of video indexing system.The information you provide will be part of aggregated data and your annotations and propositions will not be individually identified. All feedback will be confidential and will only used for research purposes.

Please Tick the relevant ☑

What is your name (Optional) _____

What is your gender Male ☐ Female ☐

What is your age? 18-25 ☐ 26-39 ☐ 40-60 ☐ Above 60 ☐

What is your highest level of education? College ☐ Undergraduate ☐ Postgraduate ☐

How do you rate your computer skills? Beginer ☐ Intermediate ☐ Advance ☐

Which browser do you use Internet Explorer ☐ Mozilla Firefox ☐ Google Chrome ☐

(Write in box)

Other _____

Please Tick the relevant ☑

	Strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
I would like to use system again	☐	☐	☐	☐	☐
I find system to be easy to use	☐	☐	☐	☐	☐
I don't think I need special training to use system.	☐	☐	☐	☐	☐
I think most people will easily use system.	☐	☐	☐	☐	☐
I feel very confident while using the system	☐	☐	☐	☐	☐

Please Tick the relevant ☑

	strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
Position of video on screen is good?	☐	☐	☐	☐	☐
Consistent use of terms throughout the system	☐	☐	☐	☐	☐
Prompt for input are relevant?	☐	☐	☐	☐	☐
Designed for all level of users	☐	☐	☐	☐	☐
The information organized on the screen is optimised?	☐	☐	☐	☐	☐

Please Tick the relevant ☑

	strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
Overall I am satisfied with how easy the system is.	☐	☐	☐	☐	☐
Error messages were properly notified?	☐	☐	☐	☐	☐
It was simple to use system.	☐	☐	☐	☐	☐
I can effectively complete my task using the system.	☐	☐	☐	☐	☐
I was able to complete my tasks in reasonable time.	☐	☐	☐	☐	☐

Please Tick the relevant ☑

	strongly agree	Agree	Not Agree / Disagree	Disagree	Strongly Disgree
The system is visually appealing.	☐	☐	☐	☐	☐
The overall organization of the system is easily understandable.	☐	☐	☐	☐	☐
System is well designed.	☐	☐	☐	☐	☐
Terminology used in the system is clear.	☐	☐	☐	☐	☐
The result met my expectation.	☐	☐	☐	☐	☐
I would like to use this system again in future.	☐	☐	☐	☐	☐
Overall, the system is easy to use.	☐	☐	☐	☐	☐

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:od="urn:schemas-microsoft-com:officedata">
<xsd:element name="dataroot">
<xsd:complexType>
<xsd:sequence>
<xsd:element ref="video" minOccurs="0" maxOccurs="unbounded"/>
</xsd:sequence>
<xsd:attribute name="generated" type="xsd:dateTime"/>
</xsd:complexType>
</xsd:element>
<xsd:element name="video">
<xsd:annotation>
<xsd:appinfo>
<od:tableProperty name="Orientation" type="2" value="0"/>
<od:tableProperty name="OrderByOn" type="1" value="0"/>
<od:tableProperty name="DefaultView" type="2" value="2"/>
<od:tableProperty name="GUID" type="9" value="TgITVKO/9kmZUqHmNsRfvA==
"/>
<od:tableProperty name="DisplayViewsOnSharePointSite" type="2" value="1"/>
<od:tableProperty name="TotalsRow" type="1" value="0"/>
<od:tableProperty name="FilterOnLoad" type="1" value="0"/>
<od:tableProperty name="OrderByOnLoad" type="1" value="1"/>
<od:tableProperty name="HideNewField" type="1" value="0"/>
<od:tableProperty name="BackTint" type="6" value="100"/>
<od:tableProperty name="BackShade" type="6" value="100"/>
<od:tableProperty name="ThemeFontIndex" type="4" value="-1"/>
<od:tableProperty name="AlternateBackThemeColorIndex" type="4" value="-1"/>
<od:tableProperty name="AlternateBackTint" type="6" value="100"/>
<od:tableProperty name="AlternateBackShade" type="6" value="100"/>
<od:tableProperty name="DatasheetGridlinesThemeColorIndex" type="4" value="-1"/>
<od:tableProperty name="DatasheetForeThemeColorIndex" type="4" value="-1"/>
</xsd:appinfo>
</xsd:annotation>
<xsd:complexType>
<xsd:sequence>
<xsd:element name="VID" minOccurs="0" od:jetType="longinteger" od:sqlSType="int" type="xsd:int">
<xsd:annotation>
<xsd:appinfo>
<od:fieldProperty name="ColumnWidth" type="3" value="-1"/>
<od:fieldProperty name="ColumnOrder" type="3" value="0"/>
<od:fieldProperty name="ColumnHidden" type="1" value="0"/>
<od:fieldProperty name="TextAlign" type="2" value="0"/>
<od:fieldProperty name="AggregateType" type="4" value="-1"/>
<od:fieldProperty name="CurrencyLCID" type="4" value="0"/>
</xsd:appinfo>
</xsd:annotation>
</xsd:element>
<xsd:element name="VName" minOccurs="0" od:jetType="text" od:sqlSType="nvarchar">
<xsd:annotation>
<xsd:appinfo>
<od:fieldProperty name="ColumnWidth" type="3" value="-1"/>
<od:fieldProperty name="ColumnOrder" type="3" value="0"/>
<od:fieldProperty name="ColumnHidden" type="1" value="0"/>
<od:fieldProperty name="TextAlign" type="2" value="0"/>
<od:fieldProperty name="AggregateType" type="4" value="-1"/>
<od:fieldProperty name="CurrencyLCID" type="4" value="0"/>
</xsd:appinfo>
</xsd:annotation>
<xsd:simpleType>
<xsd:restriction base="xsd:string">
<xsd:maxLength value="255"/>
</xsd:restriction>
</xsd:simpleType>
</xsd:element>
<xsd:element name="VLength" minOccurs="0" od:jetType="longinteger" od:sqlSType="int" type="xsd:int">
<xsd:annotation>
<xsd:appinfo>
<od:fieldProperty name="ColumnWidth" type="3" value="-1"/>
<od:fieldProperty name="ColumnOrder" type="3" value="0"/>
<od:fieldProperty name="ColumnHidden" type="1" value="0"/>
<od:fieldProperty name="TextAlign" type="2" value="0"/>
```

1

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:od="urn:schemas-microsoft-
  com:officedata">
<xsd:element name="dataroot">
<xsd:complexType>
<xsd:sequence>
<xsd:element ref="segment" minOccurs="0" maxOccurs="unbounded"/>
</xsd:sequence>
<xsd:attribute name="generated" type="xsd:dateTime"/>
</xsd:complexType>
</xsd:element>
<xsd:element name="segment">
<xsd:annotation>
<xsd:appinfo/>
</xsd:annotation>
<xsd:complexType>
<xsd:sequence>
<xsd:element name="VID" minOccurs="0" od:jetType="longinteger" od:sqlSType="int" type="xsd:
  int"/>
<xsd:element name="SNo" minOccurs="0" od:jetType="longinteger" od:sqlSType="int" type="xsd:
  int"/>
<xsd:element name="SST" minOccurs="0" od:jetType="longinteger" od:sqlSType="int" type="xsd:
  int"/>
<xsd:element name="SET" minOccurs="0" od:jetType="longinteger" od:sqlSType="int" type="xsd:
  int"/>
</xsd:sequence>
</xsd:complexType>
</xsd:element>
</xsd:schema>
```