Affect-Based Indexing and Retrieval of Multimedia Data

Ching Hau Chan

A dissertation submitted in fulfillment
of the requirements for the award of

Doctor of Philosophy (Ph.D.)

Supervisor: Dr. Gareth J.F. Jones

School of Computing

October 2006

Declaration

I hereby certify that this material, which I now submit for assessment in the programme of study leading to the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my own work.

Signed : _____

ID No. : 53149866

Date : 31/10/06

Title: Affect-based Indexing and Retrieval System of Multimedia Data

Abstract

Digital multimedia systems are creating many new opportunities for rapid access to content archives. In order to explore these collections using search, the content must be annotated with significant features. An important and often overlooked aspect of human interpretation of multimedia data is the affective dimension. The hypothesis of this thesis is that affective labels of content can be extracted automatically from within multimedia data streams, and that these can then be used for content-based retrieval and browsing. A novel system is presented for extracting affective features from video content and mapping it onto a set of keywords with predetermined emotional interpretations. These labels are then used to demonstrate affect-based retrieval on a range of feature films. Because of the subjective nature of the words people use to describe emotions, an approach towards an open vocabulary query system utilizing the electronic lexical database WordNet is also presented. This gives flexibility for search queries to be extended to include keywords without predetermined emotional interpretations using a word-similarity measure. The thesis presents the framework and design for the affect-based indexing and retrieval system along with experiments, analysis, and conclusions.

Acknowledgements

Firstly I would like to thank my supervisor, Dr. Gareth J.F. Jones. I could not have imagined having a better supervisor and mentor for my PhD, and without his kindness, knowledge, scrutiny, intuition, guidance and care over the years I would never have finished. I would like to thank my examiners, Dr. Noel E. O'Connor (internal) and Dr. Colin Johnson (external) for taking the time to read and examine the thesis as well as providing helpful feedback. Thanks to all the people in the Centre of Digital Video Processing at DCU especially Prof. Alan F. Smeaton, Dr. Hyowon Lee, and Dr. Bart Lehane for helpful guidance and advice over the years. Thanks to all the people who committed a significant amount of time to volunteer for the user experiments. I would also like to thank all the rest of the academic and support staff of the School of Computing and Dublin City University.

On a more personal note, I would like to thank my parents for their patience and active support and encouragement throughout the years. Thanks to Mabel as well for offering her support and help as well. Finally, thanks to all my friends and family for helpful advice and words of encouragement.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 1 Introduction

1.1 Background and Motivation of This Thesis

1.1.1 Explosion of Digital Video Data

We live in a world driven by technology. Rapid advances in technology have not only allowed us to travel the world and broaden our horizons but also to facilitate our freedom of expression. We record our experiences not only to remind ourselves of the past but also to share and inspire others. We create movies and TV shows to entertain and educate. Moving pictures with sound represents one of the best ways to share the experiences that we have encountered. The unique characteristic of video is its ability to convey a rich amount of semantic information through audio, visual, text, and time. Nowadays most videos are stored in digital form to prevent deterioration or loss of quality. In addition to which, it also makes replication, transfer, storage, and crucially access to the information much easier.

While not many years ago our desktop computers struggled to display compressed video, modern tiny devices such as Apple's iPod can play high-quality videos on their tiny screens. Cheap and powerful consumer devices such as digital cameras and video players make it affordable for anyone to create their own content. Handheld or portable devices in particular are seeing convergence in terms of functionality and video capture is one of them. Increasingly, mobile phones, personal digital assistants, and laptops have some sort of video capture capability. The relentless pace of increased processing power coupled with cheaper manufacturing costs has enabled more people to start recording their own digital content archives.

The pervasiveness of video capture makes it feasible to document almost every aspect of our lives no matter where we are. Research work is being carried out to try to document as much of our daily lives as possible through the use of wearable sensors capturing images and other data [Gemmell et al., 2004]. Everyone is taking advantage by recording their own content for a variety of different reasons. These include recording

interesting events, entertainment, or even for security. For example, people who keep a web log on the Internet of their daily lives or interesting events are starting to create and embed video clips into their web logs. This has led to an explosion of video data in recent years. The digital nature of these devices also makes it easy for people to upload, edit, and distribute their own videos.

1.1.2 Digital Video Delivery Through the Internet

In addition to the developments in computing devices, the introduction and rapid uptake of broadband Internet access is pushing digital video as an increasingly popular format for distributing content and information. More people are switching to broadband Internet access such as Digital Subscriber Lines (DSL) in their own households to take advantage of the increased bandwidth, as shown in Table 1-1. Even outdoors, the availability of wireless Internet connections such as Wi-Fi hotspots is making it easier to access information from the Internet.

Table 1-1. Internet access theoretical speeds.

Connection type	Bandwidth
Dial-up	Up to 56Kbit/s
ISDN	Up to 128Kbit/s
DSL (ADSL)	Up to 9Mbit/s
Wi-fi (802.11g)	Up to 54Mbit/s

As of 2005, DSL is the leading broadband platform in 28 countries, including Ireland [OECD].

Table 1-2. Sample of video format bit rates.

Video format	Bitrate
VCD	1.5 Mbit/s
DVD	5 Mbit/s
HDTV	10 Mbit/s

The increased bandwidth is cutting down the time it takes to transfer large compressed video files through the Internet, as shown in Table 1-2. The ubiquity of the Internet utilizing new technologies such as Wi-Fi makes digital video delivery very convenient. The huge number of broadband-enabled Internet users, which are increasing

every day, will want to have variety in the presentation of their information and this means demand for digital video through the Internet is set to increase as well.

Video stored in a central location somewhere can be accessed by anyone connected to the network. Continually improving video compression and streaming technologies allow us to stream high-quality videos through the network. It is now feasible to watch live news broadcasts over the Internet. As well as their traditional broadcast services, major broadcast corporations such as CNN and the BBC now broadcast their news video content over the Internet in addition to providing written text on their websites.

Besides downloading, more and more people are also uploading their own content alongside the professionals. The increased bandwidth makes it easy and convenient to upload their own videos onto Internet video hosting sites for other people to download and share.

1.1.3 Emergence of Video On-Demand Delivery

The emergence of video-centric sites such as Google Video [Google Video] and YouTube [YouTube] represents a new model in how we access video content. These sites offer a video hosting service that allows ordinary people to upload their own videos to be shared with the whole world. People can browse through thousands of videos using a text query search system similar to text-based web search engine. Although services of this type are in their infancy, they already demonstrate how the way we watch videos is shifting from pre-set schedules to a more immediate on-demand model of video delivery. On a traditional broadcast medium such as television, hard disk-based Personal Video Recorders (PVR) such as TiVo [TiVo] and ReplayTV [ReplayTV] can be purchased to automatically record and sort videos for on-demand viewing at the user's leisure. Even on modern third-generation (3G) mobile phone networks, short downloadable news videos or sports highlights are being developed as a viable revenue stream.

The explosion of so much video data, when coupled with an increasing appetite for on-demand delivery through the Internet, means efficient and powerful multimedia system solutions capable of indexing and retrieval of content relevant to the user is

needed critically. The next section highlights some of the issues of video retrieval that are being addressed in current research and introduces the topic of affect in video and its potential application in video retrieval.

1.2 Significance of Research Work in the Thesis

1.2.1 Information Overload

As described above since it is easy now for anyone to make and distribute digital videos, the amount of professional and personal video data is set to increase dramatically over the coming years. With such a huge volume of data, manually searching for a piece of video from within a large database, which is already a time-consuming and tedious task, is set to become entirely impractical. Due to its temporal nature, even skimming through a single piece of video quickly to look for something specific can be taxing to the viewer due to the amount of mental concentration needed.

But of course, not all pieces of video are annotated, which makes the process of identifying what videos contain even harder. The richness of semantic information contained in video also makes it difficult to properly describe all events or information contained within it even if manual effort is made to annotate it. In some cases, the person searching might not even know if a video matching their requirements exists, in which case they need to be presented with the closest possible match. In addition, different people could interpret a video event differently which makes the annotation and retrieval task even harder.

In order to most effectively exploit the huge volumes of video data, users require annotation tools for selection and classification of relevant content. This has led to a significant growth in recent years in research into content-based digital video search technology or content-based video retrieval (CBVR) aimed at helping users to locate and retrieve relevant video from potentially very large databases.

New automated annotation methods need to be investigated to bring out as much information as possible to annotate and retrieve video in order to fulfill the growing information needs of video users.

1.2.2 Content-Based Video Retrieval on the Affective Level

Not much related work currently exists in indexing and classifying affective features for multimedia content analysis. The emerging interdisciplinary field of *Affective Computing* is significantly improving our understanding of the relationship of human-computer interaction and human emotional states to which intelligent machines can react accordingly to provide better results to our information needs [Picard and Cosier, 1997]. According to Eakins [1996] and Hanjalic and Xu [2003], video content perception, analysis and retrieval can be differentiated into 3 levels as shown in Figure 1-1. Video can be retrieved at the lowest level (feature level) using primitive features such as shot cuts and camera motion. This is then followed by logical features (cognitive level) that are higher level and involves some degree of logical inference about its identity. The highest level (affective level) contains the most abstract features that involve some degree of subjectivity. These levels are discussed in more detail in Section 3.4 (The Three Levels of Content Retrieval).

Existing work on multimedia content analysis has concentrated largely on recognition of the objective features, hence improving efficiency and reliability at the feature and cognitive level [Hauptmann and Christel, 2004] for content-based video retrieval systems. Examples of features at the feature and cognitive level is shown in Figure 1-1. Such work includes scene segmentation [Zhang and Jay Kuo, 2001] and [Zhai et al., 2004], object detection [Browne and Smeaton, 2004], and sports highlight detection [Sadlier and O'Connor, 2005]. These features allow users to make queries that are clearly defined and unambiguous. A number of current large-scale content-based video retrieval systems based on using these features have been developed as described by Smeaton, [2004]. The feature indexing tools within these systems rely on extraction of low-level feature analysis such as audio energy, pitch or the colour distribution and texture in the video.

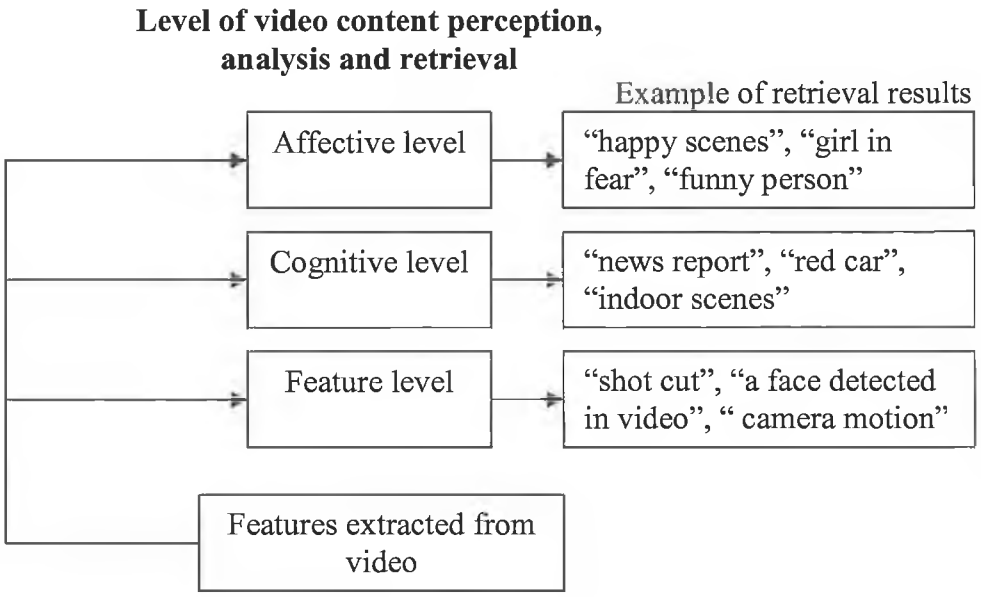


Figure 1-1. Overview of 3 different levels of video content perception, analysis and retrieval, taken from [Hanjalic and Xu, 2003].

Affective features contain information about emotive elements which are aimed at invoking certain emotional states that humans can naturally identify. Therefore detecting affective features is not about detecting the viewer's emotional state, but the emotion that a director of a film is seeking to create in the viewers. Further details and a review of affective computing and its key concepts are presented in chapter 4 (Affective Computing). This review highlights new and exciting applications for human-computer interaction such as expressive avatars and better human communication and emotion-aware intelligent agents. It is clear that the affective dimension is an important natural component of human classification and retrieval of information as it offers a new approach to describe or analyze information as well as offering users a new modality of interaction with video data [Picard and Cosier, 1997]. Given the demands of CBVR systems and the very limited existing work, there is a pressing need for more investigation into video retrieval applications at the affective level.

For example, extraction of sports highlights using objective measures is an important application of video analysis. There are many approaches based on audio classification, video feature extraction and highlight modeling, or a combination of these

features. One example by Rui et al. [2000] presented an audio-based method for detecting highlights from baseball videos. Highlights are detected by tracking audio features of speech by the commentator and detecting the sound of a baseball hit. Another example of sports video highlights extraction is by Sadlier and O'Connor [2005]. Their approach detects highlights from soccer videos by tracking features from both audio and video sources. Using feature detectors for crowd image, speech, on-screen graphics, and motion they are able to detect highlights as well as to summarize the activity of a soccer video.

While this allows us to monitor the activity in a sports event, it doesn't allow us to track the emotional aspect of sports. Affective computing techniques could be applied to sports video summarization to extend its capabilities beyond numbers and figures. Being able to monitor the emotional experience using analysis of individual players' affective features such as facial expressions or body language or affective states of crowd behaviour might be useful to complement feature detectors that are looking for scenes of violence or a foul.

Work in this area is relatively new and therefore has potential for many applications. This thesis investigates the application of affective computing techniques in CBVR to develop a new model of user interaction with CBVR systems. Emotions in video, which previously have been largely ignored, hold much information that can be naturally interpreted by a human viewer. Therefore it is also natural that humans express their information need in terms of emotional content. Current work in CBVR is focused on classifying and retrieving semantic units such as events and objects that are clearly defined. Emotions however, are not clearly defined, but essential as they form a large part of video content. Through the application of affective computing, this thesis hopes to pioneer work in this direction to enable users of CBVR systems to express their information needs and locate scenes taking into account affective dimensions.

1.2.3 Ambiguity and Incompleteness of Video Annotation and Search

People looking for a piece of video from a video archive typically do two kinds of search, they either have in mind a specific piece of video they are looking for (a "known item" search) or a broader search to look for something that fits some general requirements. A

CBVR system therefore, has to be able to retrieve a requested video almost instantaneously, recommend other videos of similar criteria if an item the user has in mind cannot be found.

Taking the earlier example of popular video hosting sites on the web such as [YouTube] and [Google Video], these currently rely on manual text descriptions provided by human users. Uploaded videos are annotated with text descriptions (called tags or metadata such as "holiday video in Paris" or "cat climbing a tree") which are written by the user who uploaded the video. Aside from browsing through the video itself, people use these tags to help them find what they're looking for. These tags although useful, usually end up losing a lot of contextual information and consistency as it is also impractical to expect users to provide rich detailed annotations. A video with a vague tag like "skateboard" does not specify whether it is a shot of a skateboard or a skateboarding event or the person riding the skateboard.

While this approach might work well for a video hosting site with many human users uploading and annotating their own favourite videos to share, a digital video retrieval system usually has no such luxury, and needs to automatically generate consistent and useful metadata for a large volume of video as shown in Figure 1-2.

Figure 1-2. Manually annotating videos with metadata is time-consuming and needs to be automated to support diverse user queries in many situations.

1.2.4 Video Indexing with Emotion Labels

The annotation task becomes harder when more abstract descriptions such as affective labels are needed. In this situation videos depicting a scene or a mood need to be labeled

with descriptions that are abstract such as "blue" or "sad". The tedious nature of manually thinking up a set of tags and the time required to do it can also discourage people from finding the ideal description. The subjective nature of such labels also means that manual annotation is more likely to be more difficult than with objective features. A particular problem is likely to be the inconsistent labels that would be manually assigned when different users interpreted emotions differently. Even individual users find it hard to behave consistently for different videos.

Current work into video retrieval systems is focused on the detection of clearly defined semantic objects. These objects describe concepts such as people, face, cars, trees or buildings that appear in a piece of video. Users of such system can make specific queries such as "Jay Leno" or "red car". Videos have emotive elements, but research has not previously been pursued to allow users to use them in their queries. Current video retrieval systems don't accommodate queries such as "video of a happy Jay Leno in a red car" as there is no clear definition for the word "happy". Users would need to browse all videos containing "red car" and "Jay Leno" in order to find it. A mechanism for distinguishing between "happy" and other video segments would be useful. This makes the emotive elements of a video provide a helpful context to a piece of video, as well as forming part of the solution to bridging the semantic gap between user needs and annotation features of CBVR systems (Section 3.3 Bridging the Semantic Gap with Affective Computing). The emotional content in a video can potentially be inferred using extraction of features associated with its affective dimension. This thesis investigates and presents a novel affect-based video retrieval system that detects affective states in video for annotating them with emotion text labels, and captures its use on affect-based retrieval of films based on the labels.

1.3 Contributions

Hanjalic and Xu, [2003] presented a method to detect and model affective features for audio-visual data. Using video and audio low-level features such as camera motion, shot cuts, audio energy and audio pitch, they demonstrated an approach which combines these features to detect affective features that are said to influence a viewer's emotions. Their

work stopped at only presenting a method for modeling these affective features. A survey of related published work reveals that no previous work has been done to solve the problem of translating these features into useable emotional labels and applying it in a video retrieval system.

The key contribution made in this thesis is the mapping of these affective features onto emotional labels to enable content-based video retrieval. The unique feature of such a novel approach compared to other related work is that it does not rigidly categorize scenes into only a few categories of emotion. Such rigidity would mean that any scene is always associated with only one broad emotion or the other. The approach presented here allows a far wider range of emotions to be described, with a finer granularity and greater flexibility of search and retrieval.

In addition, the work presented here also recognizes the fact that emotions can overlap and minor emotions that might exist in the data are not removed completely, giving a more accurate picture of the development of emotions as well avoiding the removal of significant related features of information about the content of the data.

This is further extended by using classic text retrieval techniques to sort and rank videos according to their emotional labels and implementing an open-vocabulary query system to enable users to use emotional labels that are not covered in the basic system, thus bridging a semantic gap between system-detected features and user expression of information requirements. Therefore the bulk of this thesis' novelty is the design of an affect-based video indexing and retrieval system by extending current methods of affective feature extraction, in representation and labeling videos with emotional labels and in allowing users to query and retrieve video segments from a suitably large corpus of Hollywood films. The thesis also presents the necessary literary background, issues, and challenges facing the design of an affect-based video indexing and retrieval system along with experiments, analysis, and conclusions. No published work has been found that discusses the design of such an affect-based system running on a substantial amount of video data as presented here, therefore these elements can be considered to be another key novelty of this thesis.

The experimental investigation forms the other significant part of this thesis' novelty. No previous and similar work had been done to retrieve video using only a

textual description of its emotional content, so there were no guidelines and baseline results that could be used as reference. The methodology and results presented in this thesis is therefore a novel set of experimental data that can be referenced for future work.

1.4 Thesis Overview

This thesis is organized as follows: Chapter 2 introduces some relevant concepts of how digital video and audio are represented and stored, Chapter 3 introduces and discusses related issues regarding content-based video retrieval systems that are relevant to this thesis, Chapter 4 provides justifications for applying affective computing techniques into content-based video retrieval, Chapter 5 discusses classic information retrieval techniques used in the affect-based video indexing and retrieval system, and an introduction to WordNet and relatedness measures used here for open-vocabulary query mapping, Chapter 6 presents the design of the novel affect-based video indexing and retrieval system, Chapter 7 presents an experimental investigation of the system using a film collection, and finally Chapter 8 provides conclusions and future work.

Chapter 2 Digital Video

In order to understand how captured video can be processed to generate useful metadata, some knowledge of selected fundamental concepts of digital video and audio along with a review on a popular encoding format is necessary. This chapter introduces some fundamental concepts of digital video and audio relevant to the development of content-based video retrieval (CBVR) systems.

2.1 Digital Video Concepts

2.1.1 Introduction

Video has become an integral part of our daily lives, for example on television, in movies, as home videos, on digital billboards, and increasingly today even on mobile devices. Video is a constant source of knowledge, inspiration, and entertainment, driven by technology and commerce.

Video is a communication medium that consists of a series of still images synchronized to audio. The series of still images or frames are displayed in quick succession to achieve the illusion of movement. For video information retrieval (IR), the following terms can be defined:

- Frames – a frame is a single image in a video stream
- Shots – a shot is a single continuous recording of a camera over a period of time
- Shot cut – a shot cut is said to occur between two shots, this is usually signified by a sudden or gradual transition from one shot to the other
- Scene – a scene is a group of semantically related shots
- Keyframe – a keyframe is a single chosen frame to represent a shot or scene

Video IR systems typically allow users to search for relevant shots which are then represented to the user as a single keyframe. Shots are identified in the video by

automatic detection of shot cuts. Much more difficult is the automatic detection or even the definition of, scene changes. Users are often shown potentially relevant shots with keyframes representing preceding and following ones so that the user can infer the content and identify scene boundaries for themselves [Smeaton et al., 2004].

2.1.2 Digital and Analog Video

Video can be represented either by using a digital or analog representation of the video signal. Before the introduction of digital video systems, all video signals were produced and stored in an analog format that was typically recorded onto magnetic tapes. Analog signals encode picture and sound information by varying the amplitude and frequencies of a signal. In analog television, different encoding schemes are used in different regions of the world, culminating in 3 major standards known as PAL (Phase Alternating Line), NTSC (National Television Systems Committee) and SECAM (Séquentiel couleur à mémoire, French for "sequential colour with memory"). The resultant analog signal from these encoding schemes can then be modulated onto a radio-frequency carrier and broadcast through an antenna. Analog television signals are typically stored on a magnetic tape format such as Video Home System (VHS).

As the name suggests, digital video is video information that is captured and stored in a digital format. The video information is converted into a series of electrical pulses in binary numeric form, represented by a series of 0s (pulse absent) and 1s (pulse present). Nowadays, most video is produced in digital form. Digital video has typically been recorded onto magnetic tapes. The recorded signal is then broadcast through a digital television system or distributed on optical discs such as Digital Versatile Disc (DVD). Recent digital video recorders are also introducing the option of recording directly onto a digital format such as optical discs or hard drive media. The remainder of this subsection introduces several important concepts in video:

- **Frame Rate**

The series of still images in a video are called frames. The frames are displayed according to a set frame rate to maintain the illusion of movement. Frame rate in video refers to the number of frames per second (fps) that dictates how rapidly the frames are to be displayed. Natural and realistic movement usually requires a minimum frame rate of around 25 fps, therefore television features a frame rate of either 25 fps (PAL) or 30 fps (NTSC). Digital encoding of video, such as in Moving Picture Experts Group (MPEG) encoding, has also adopted 25 fps as the default setting. MPEG is discussed in Section 2.3 (A Review on MPEG Compression).

- **Video Resolution**

The size of a video frame is measured in pixels (picture element) for digital video or horizontal scan lines for analog video. This is the spatial resolution of the video. For standard-definition television (SDTV), the resolution is 480 horizontal lines for NTSC televisions, while PAL television has a resolution of 576 horizontal lines. Resolution in digital video can be expressed as "640x480 pixels", giving a total of 307200 pixels. Currently emerging high-defmition television (HDTV) has a resolution of 1920x1080 pixels, producing extreme clarity and detail. It is also common to find digital video encoded at lesser resolutions such as 320x200, used for streaming on the Internet due to its limited bandwidth.

- **Aspect Ratio**

Aspect ratio describes the shape of the video. Standard television displays use an aspect ratio of 4:3 or 1:33:1, which gives a square shape, while high-definition television and cinemas use an aspect ratio of 16:9 or 1:78:1 sometimes called "widescreen" due to its rectangular shape. Although digital video signals typically conform to these ratios, they are not limited by it.

- **Bits per Pixel**

The number of colours that can be represented by a pixel depends on the number of bits per pixel (bpp), also known as colour depth. Each pixel holds colour information, the higher number of bits assigned to each pixel, the broader the range of colours that can be displayed. For example, a 1-bit colour depth (also known as monochrome) gives $2^1 = 2$ colours, an 8-bit colour depth (also known as VGA or video graphics array) gives $2^8 = 256$ colours, and a 24-bit colour depth (also known as Truecolor) gives $2^{24} = 16.7$ million colours

- **File Size and Compression**

One key issue of digital video is the file size. Because frames of video need to be displayed at a rate of 25 frames per second, even without including the accompanying audio, the file size can become very large very quickly. Therefore a compression scheme is almost always used to store digital video files. The most popular standards for video compression are the MPEG suit of standards. The MPEG-1 compression scheme is described in detail in Section 2.4 (MPEG-1 Video). The following is the formula for calculating an uncompressed video file size without audio,

Uncompressed video size = (no. of bits) = frame height (no. of pixels) × frame width (no. of pixels) × colour bit depth (bits per pixel) × frames per second (fps) × video length (no. of seconds)

Therefore, calculating the size of a normal 1 second SDTV video clip,

184320000 bits or 23 Mbytes = 640 h × 480 w × 24 bpp × 25 fps × 1 second

For compressed video and audio files, it is convenient to describe their file size in terms of the bit rate information delivery. Bit rate is expressed in bits per second (bps) or

Megabits per second (Mbps). Higher bit rates usually mean better quality. Table 2-1 compares the bit rates for different video type:

Table 2-1. Bit rates of different video types

Video types	Bit rate
Video conferencing	128-384 Kbps
Video CD (VCD)	1 Mbps
DVD	5 Mbps
HDTV	15Mbps

2.1.3 Digital Video Navigation and Browsing

Video is a temporal medium, and the conventional way to browse video has always been based on a temporal model such as found on VCR and DVD players. The "play", "rewind", "fast-forward", and "pause" concepts are intuitive and easy to understand. Coupled with a timeline to indicate the temporal location in which video is being played back, a user can jump to any point of video quite easily. Digital video accessed through a computer uses the same conventions by mimicking a virtual player for users to interact with. Thus, while digital video allows immediate access to any point in the video, concepts of browsing with users are based on the familiar metaphor of a physical video tape player. This restricts the user from using the immediate access feature of digital video more effectively.

Figure 2-1. Example of a video media player with a timeline.

As can be seen in Figure 2-1, a digital media player delivers video on a main window and allows the viewer to navigate through the video with various buttons. These include the pause button (the biggest icon at the bottom left), stop button (right of the pause button), forward and rewind buttons (right of the stop button), volume and adjustment slider (right of forward button) and the long green bar above it is the timeline slider, allowing the user to quickly skip to any point in the video (in this example this is set to the middle of the video).

2.1.4 Colour Space Representation in Digital Video

One important aspect of digital video is how colour is represented. There are numerous colour representations, also called colour space in which video is encoded. The values used to store colour information associated with the pixels take up the bulk of the video data. Each colour space has unique advantages, therefore the data is converted from one colour space to the other and back again depending on the video needs or tasks. The following subsection reviews the colour spaces used in this thesis to achieve video feature extraction and analysis.

- **RGB Colour Space**

Digital video on computer monitors typically uses the RGB colour space to represent colours, although there are other colour spaces which can be used for this. RGB, which stands for red, green and blue, is a convenient colour space to use as it mimics the way the human visual system works. It also happens to be simple to understand. An RGB signal or image is stored in its separate red, green and blue channels, which simply describe how much red, green and blue colour information there is in each pixel. Pixels vary between minimum (black) to maximum (full intensity). The number of gradations between minimum and maximum depends on the colour depth (bits per pixel). Therefore the more bits per pixel, the more distinct colours can be represented.

Figure 2-2. An image represented in the 3 RGB channels. Taken from [Wikipedia].

For example, a Truecolor (24-bit) image such as shown in Figure 2-2, would have 3 integers representing red, green and blue intensities each within the range 0 to 255 ($3x2^8$ bits) at each pixel location. A (0,0,0) value refer to the colour black, (255,0,0) to the colour red, and (255,255,0) to the colour yellow.

- **YUV and YCbCr Colour Space**

The YUV colour space defines the colour space in terms of one luminance component and two chrominance components. The PAL and NTSC television standards use this colour model, while SECAM uses a variant of this called YDbDr to encode colour information. Y is the brightness or luma channel, and U and V are the colour difference channels, which correspond to Cb and Cr as well. This colour space has an advantage over the RGB colour space. Historically, in analog television, the YUV colour space was used because it remains compatible with black and white analog television. The Y channel retains all data that is needed to display a monochrome image while the U and V channels can be discarded. For a colour television, all three channels are used and the RGB information can be decoded and colour displayed. This produces a signal that is compatible for both monochrome and colour television. As can be seen in Figure 2-3, the original image on the left is separated into the Y, U and V channel.

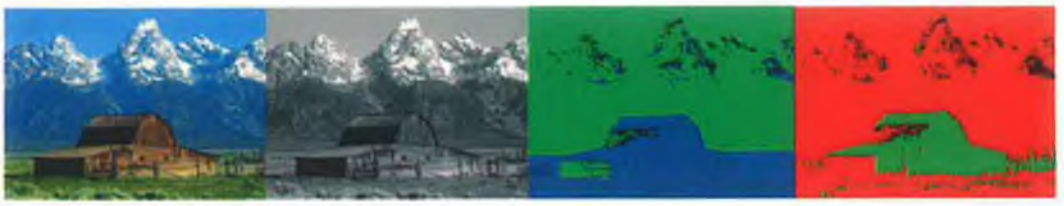


Figure 2-3. An image with its 3 YUV channels, note that the Y channel can be used as a monochrome image. Taken from [Wikipedia].

The YUV colour space can be derived from a RGB source through the following formula,

$$Y = 0.299 \times R + 0.587 \times G + 0.114 \times B$$
$$U = -0.147 \times R - 0.289 \times G + 0.436 \times B$$
$$V = 0.615 \times R - 0.515 \times G - 0.100 \times B$$

In digital video, the YUV colour space can be mistaken for the YCbCr colour space, although both are sometimes used interchangeably. An analog encoding scheme produces the YUV colour space, while a digital encoding scheme produces the YCbCr colour space. It is fundamentally the same except for the scale factors on the chroma components that are different (U, V and Cb, Cr). Therefore, digital video files such as those encoded in MPEG are encoded in the YCbCr colour space.

Similar to YUV, the YCbCr colour space can be derived from an RGB source,

$$Y = (0.299 \times R + 0.587 \times G + 0.144 \times B) - 128$$
$$Cb = 0.433 \times (B-Y)$$
$$Cr = 0.877 \times (R-Y)$$

This colour space is still used in the digital domain because of its compression advantages which will be discussed further in the MPEG Section 2.4 (MPEG-1 Video). YCbCr can be designated a ratio (4:n:n). This describes the ratio between the Y, Cb and

Cr channels. While RGB has the full ratio of 4:4:4, other colour models can have different ratios. The 4 represents a sampling rate of 13.5 MHz, which is the standard frequency, defined by ITU-R BT.601 for digitizing analog NTSC, PAL and SECAM signals. The 4:2:2 colour sampling ratio is often used for high quality sampling, while 4:2:0 is used in MPEG compression. The reduction from 4:4:4 in RGB to 4:2:0 in YCbCr without much loss in quality is also known as chroma subsampling. Chroma subsampling works by sampling the 2 channels of chroma information at a lower rate than the luminance channel. This is because human eyes are less sensitive to colour than luminance, therefore less colour information can be encoded without significant loss of quality, taking up less bandwidth.

- **YIQ Colour Space**

The YIQ colour space was previously used in the NTSC television standard. I stands for in-phase, while Q stands for quadrature, with the Y representing luminance information. This colour space is similar to the YUV model in that I and Q hold chrominance information such as in U and V. The key difference is that the I and Q values are derived differently, in effect giving more emphasis to Q than I. This is because YIQ was designed with human colour-response characteristics in mind. The eye is more sensitive to changes in the orange to blue range (I) than purple to green range (Q). The high cost of implementing I and Q decoding on television has meant that YIQ was dropped in favour of YUV in the NTSC standard. The YIQ colour space can be calculated from an RGB source as follows,

$$Y = 0.299 \times R + 0.587 \times G + 0.114 \times B$$
$$I = 0.595 \times R - 0.274 \times G - 0.321 \times B$$
$$Q = 0.211 \times R - 0.522 \times G + 0.311 \times B$$

- **HSV Colour Space**

The HSV colour space is a nonlinear transformation of the RGB colour space, also known as the HSB colour space. The HSV defines the colour space in terms of 3 components:

- Hue – the colour type (such as red, blue, or yellow)
    - The value ranges from 0 to 360, cycling through all colours
- Saturation – the "vibrancy" or purity of the colour
    - The value ranges from 0-100%
- Value/Brightness – the brightness of the colour
    - The value ranges from 0-100%

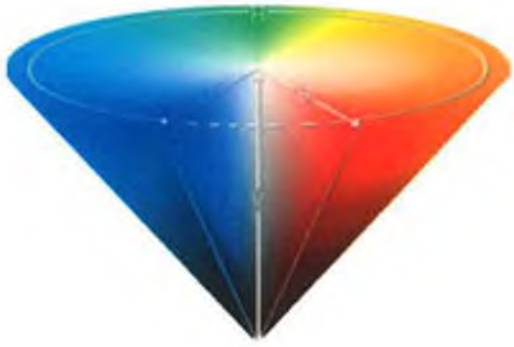


Figure 2-4. Conical representation of the HSV colour model. Taken from [Wikipedia]

As shown in Figure 2-4, the HSV colour model can be represented as a cone with H the circular circumference, S represents as radius of the cone, and V the distance from the circular area to the pointed end of the cone. The HSV colour space can be derived from the RGB colour space. Where H values ranges from 0 to 360, and S,V,R,G,B values range from 0 to 1, let MAX equal the maximum of the (R,G,B) values, and MIN equal the minimum of those values,

$$H = \text{undefined}, \qquad \text{if MAX} = \text{MIN},$$

$$H = 60 \times \frac{G - B}{\text{MAX-MIN}} + 0, \qquad \text{if MAX} = R \text{ and } G \geq B,$$

$$H = 60 \times \frac{G - B}{\text{MAX-MIN}} + 360, \qquad \text{if MAX} = R \text{ and } G < B,$$

$$H = 60 \times \frac{B - R}{\text{MAX-MIN}} + 120, \qquad \text{if MAX} = G,$$

$$H = 60 \times \frac{R - G}{\text{MAX-MIN}} + 360, \qquad \text{if MAX} = B,$$

$$S = 0, \qquad \text{if MAX} = 0,$$

$$S = 1 - \frac{\text{MIN}}{\text{MAX}} \qquad \text{if otherwise,}$$

$$V = \text{MAX}$$

While the RGB colour space is closely related to how the human visual perception works, the HSV colour space encapsulates information about a colour that is easier for us to understand and define colours in the HSV colour space than RGB or YUV. The affect-based indexing and retrieval system presented in this thesis uses this colour space because we can tell what colour is contained in an image, how dark or light it is, and how vibrant the colour is. This makes it easier to assign a subjective meaning such as tone or mood to a specific colour in terms of how dark/light and vibrant it is, thus allowing colour to be used as a feature to model affective features.

2.2 Digital Audio Concepts

2.2.1 Introduction

Digital audio comprises sound recordings of speech, music or other sounds. Sound waves propagate through air or other media and can be described as an analog signal that changes in amplitude and frequency. Digital audio signals are basically discrete-time signals, defined at equally spaced discrete values of time. Consequently, a discrete-time signal can be represented as a sequence of integers in the range from $-\infty$ to $+\infty$. However, this is in reality limited by the resolution of the samples.

Digital audio signals are represented as sequences of numbers called samples as shown in Figure 2-5. An individual sample value is denoted as x(n). An example of a digital audio signal with real-valued samples is given by,

$$x(n) = \{\ldots, 0.95, -0.2, 2.17, 1.1, \ldots\}$$

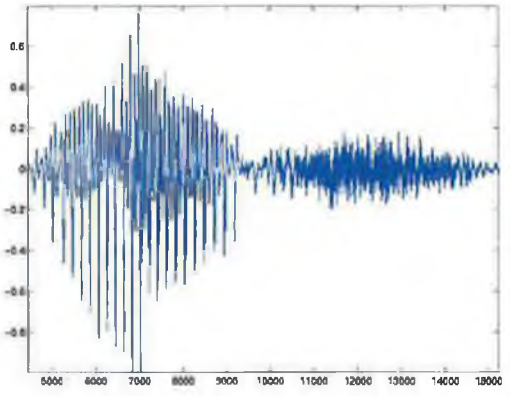


Figure 2-5. An audio signal plotted against time shows the fluctuations of a sound wave.

2.2.2 Resolution and Channel of Digital Audio

At each sample point, a value is stored to represent the amplitude of the signal. The range or resolution depends on how many bits are used to represent each sample. The more bits assigned, the more gradations there are to represent a signal, and hence the better the quality of the audio. For example, 16 bits in binary can represent $2^{16} = 65536$ values. This gives 16-bit digitized audio a range of -32768 to 32767 values centered on zero. 16-bit typically has enough resolution to sample and produce high quality audio, although it can be increased or decreased according to needs. Most digital audio uses two separate audio channels to store and playback to create a stereo effect.

2.2.3 Pulse-Code Modulation (PCM)

Similar to video, uncompressed audio files can take up a lot of space. Digitized audio can be stored in many different formats. One of the most well known formats is the WAV file format. As a Microsoft and IBM audio file format, most personal computers today are equipped for this type of file playback. Uncompressed audio is stored under a pulse-code modulation (PCM) coding in a WAV file. Uncompressed audio in this format is easy to access and manipulate.

Pulse-code modulation represents an audio signal by sampling the amplitude or magnitude at set intervals which is then quantized into a digital code. This gives an approximation of the analog signal as shown in Figure 2-6. Because PCM is a lossless format, audio WAV files encoded with PCM are ideal for editing and manipulation.

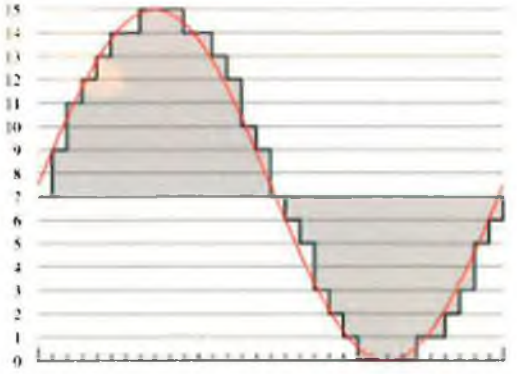


Figure 2-6. Sampling an analog signal (in red). Taken from [Wikipedia]

2.2.4 Sampling Rate and Nyquist-Shannon Sampling Theorem

The spacing between two consecutive samples is called the sampling rate or sampling period. The reciprocal of the sampling interval T, denoted as F_T, is called the sampling frequency. The unit of sampling frequency is cycles per second, or Hertz (Hz), if the sampling period is in seconds.

$$F_T = \frac{1}{T}$$

PCM represents an audio waveform by measuring the amplitude of an audio signal at set intervals. These intervals are determined by the sampling rate.

One vital consideration concerning sampling rate is the Nyquist-Shannon Sampling Theorem. The theorem states that to represent digitally a signal containing frequency component up to X Hz, it is necessary to use a sampling rate of at least 2X samples per second [Nyquist, 1928].

This means that in order to create a digital signal that contains all audible frequencies up to 20000 Hz, the sampling rate must be at least 40000 Hz per second in order to avoid a phenomenon known as foldover or aliasing. If the sampling rate falls below 40000 Hz, then the digital signal would not accurately represent the original signal.

Professional musicians store and manipulate their work in audio CD recordings sampled at a rate of 44100 Hertz (44.1 KHz). The frequency-response of human hearing generally falls between 20 Hz and 20 KHz, therefore according to the theorem, about 44.1 KHz is generally sufficient to describe the range of human hearing.

2.2.5 Representation of Digital Audio

Analog and digital signal processing can occur in the time domain. This is usually concerned with some value of the signal versus time. This is how the signals themselves

are generated, one value after another sequentially as they are generated by the source. Commonly when plotting out the shape of the audio signal, the y-axis in the time domain is the amplitude while the x-axis is the time plot, as shown in Figure 2-7. This gives us a detailed view of the changes in amplitude where information and patterns can be identified.

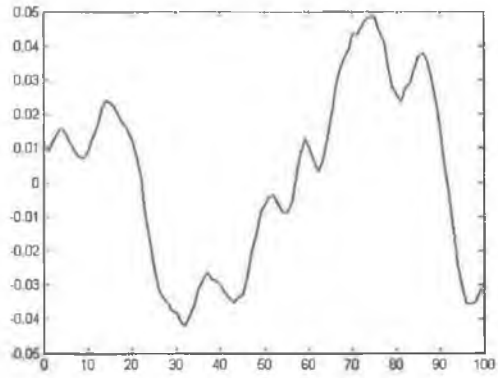


Figure 2-7. A plot of an audio signal in the time domain reveals amplitude changes.

However, just looking at the time domain is sometimes not enough as it is difficult to see what frequencies the signal contains. The signal can be transformed into the frequency domain (an example is shown in Figure 2-8) to reveal more information about its contents. In this domain, the x-axis represents frequency while the amplitude is plotted on the y-axis. Although the plot in the time domain can be directly created by plotting the amplitude at set intervals of time, the frequency domain is calculated from the time domain. The French mathematician Jean Baptiste Fourier defined a useful theory and transformation, which can be used to convert the time domain to the frequency domain and back. [Cooley and Tukey, 1965] introduced an efficient algorithm to convert signals from the time domain to the frequency domain and back called the fast Fourier transform (FFT).

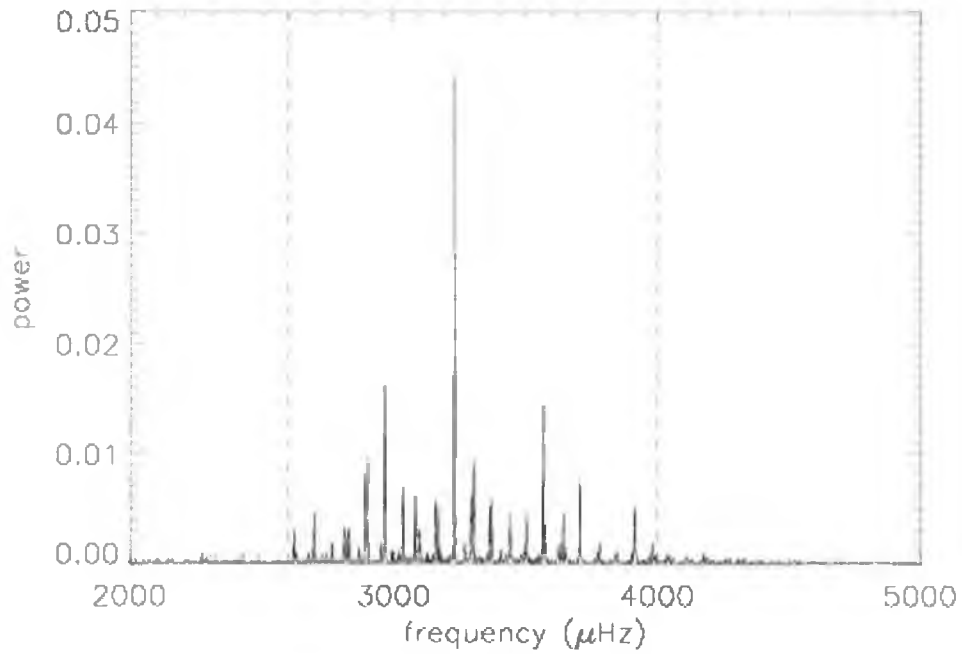


Figure 2-8. A frequency domain representation of an audio signal shows the concentration of different frequencies.

2.2.6 Fast Fourier Transform

The Fourier transform has long been used for characterizing linear systems and for identifying the frequency components making up a continuous waveform. When a waveform is sampled on a digital computer, the discrete Fourier transform (DFT) is used.

The Fourier transform pair for continuous signals can be written in the form,

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft}\,dt$$

$$X(t) = \int_{-\infty}^{\infty} x(f)e^{i2\pi ft}\,dt$$

for $-\infty < f < \infty$, $-\infty < t < \infty$, and $i = \sqrt{-1}$. X(f) and x(f) represents the frequency-domain function, X(t) and x(t) represents the continuous time function. The analogous discrete Fourier transform pair or DFT that applies to sampled versions of these functions can be written in the form,

$$X(f) = \sum_{k=0}^{N-1} x(t)e^{-i2\pi ft/N}$$

$$X(t) = \frac{1}{N} \sum_{k=0}^{N-1} x(f)e^{i2\pi ft/N}$$

The FFT is simply an efficient method for computing the DFT. It expresses the DFT of an arbitrary composite size $n = N_1N_2$ in terms of smaller DFTs of sizes N_1 and N_2, recursively to reduce computation time. The FFT can be used in place of the continuous Fourier transform only to the extent that the DFT can, but with a substantial reduction in computer processing time.

2.2.7 Audio formats

A compression scheme is often used to reduce the file size of digital audio files. There are many audio formats such as AU, VOX, AIFF, and RealAudio that use different types of compression schemes. The most popular audio format of recent years is the MPEG-1 Audio Layer 3 (MP3) format. This lossy compression format utilizes psychoacoustic models to discard information that is less audible to human hearing. This is similar in principle to chroma subsampling as described in the previous section. This compression scheme is typically able to achieve high levels of compression (such as 20:1 compared to uncompressed formats) without much loss in quality. MP3 is described in detail with other aspects of MPEG compression in the next section.

2.3 A Review of MPEG Compression

The MPEG compression standard is one of the most popular standards widely used to encode digital video and audio. The MPEG standard allows high compression rates for video and audio without a significant decrease in perceived quality. The MPEG compression standard is a result of collaborative work done by an international committee comprised of experts from the Moving Picture Expert Group (MPEG) set up by the International Standards Organization (ISO) in 1988. These standards were designed to tackle challenges facing digital video in applications such as videoconferencing and broadcasting. MPEG has standardized the following compression video and audio compression formats over the previous years [MPEG]:

- MPEG-1 – the initial video and audio compression standard (similar in design to ITU-T H.261) which includes the popular MPEG-1 Audio Layer 3 (MP3) format. This standard was designed to achieve full motion video at a rate of 1.5 Mbps, which was the transmission rate over a network at the time of its design as well as the transfer rate of CD. MPEG-1 was used for Video CD, which became the most popular distribution format throughout Asia when it was introduced.
- MPEG-2 – the standard for broadcast-quality video and audio designed for 1.5 Mbps to 15 Mbps data rates, adding support for interlaced video. MPEG-2 is used for digital television broadcast such as HDTV and DVD.
- MPEG-4 – this standard, absorbed previous MPEG standards, with enhanced features such as support for Digital Rights Management (DRM), Advanced Video Coding (AVC), technically similar to ITU-T H.264), Advanced Audio Codec (AAC), and Virtual Reality Modeling Language(VRML) support for 3D rendering.
- MPEG-7 – not a compression standard but a multimedia content description standard to store video metadata using the Extensible Markup Language (XML) to tag events and time codes occurring in a video. This standard is designed to provide a framework for describing multimedia content to streamline content manipulation, filtering, and personalization applications.

2.4 MPEG-1 Video

2.4.1 Introduction

MPEG-1 is a standard in five parts, this chapter presents the video encoding part, ISO/IEC 11172-2. Further details can be found at [MPEG]. MPEG-1 video compression achieves its high compression ratio of around 20:1 by cutting out data redundancy. Video is composed of a series of still frames that are played back rapidly to produce the illusion of movement. However, because of the high rates required to maintain a smooth movement such as 25 fps, the difference between each pair of frames is usually quite minimal. An uncompressed format would store all the frames in their complete form, therefore producing extremely large files.

In general an MPEG encoder exploits this by calculating and storing the difference between frames. The difference is based on motion estimation, in which the movement of blocks of pixels from one frame to the next is calculated. The displacements of these blocks of pixels (macroblock) are called motion vectors. Storing just the difference does not generate the complete image therefore a reference frame (Intra frames or I-type) is fully stored periodically so the differences can be reconstructed (motion estimation) into an image by the decoder. The MPEG-1 bit stream is layered as shown in Figure 2-9 and its explanation in Table 2-2.

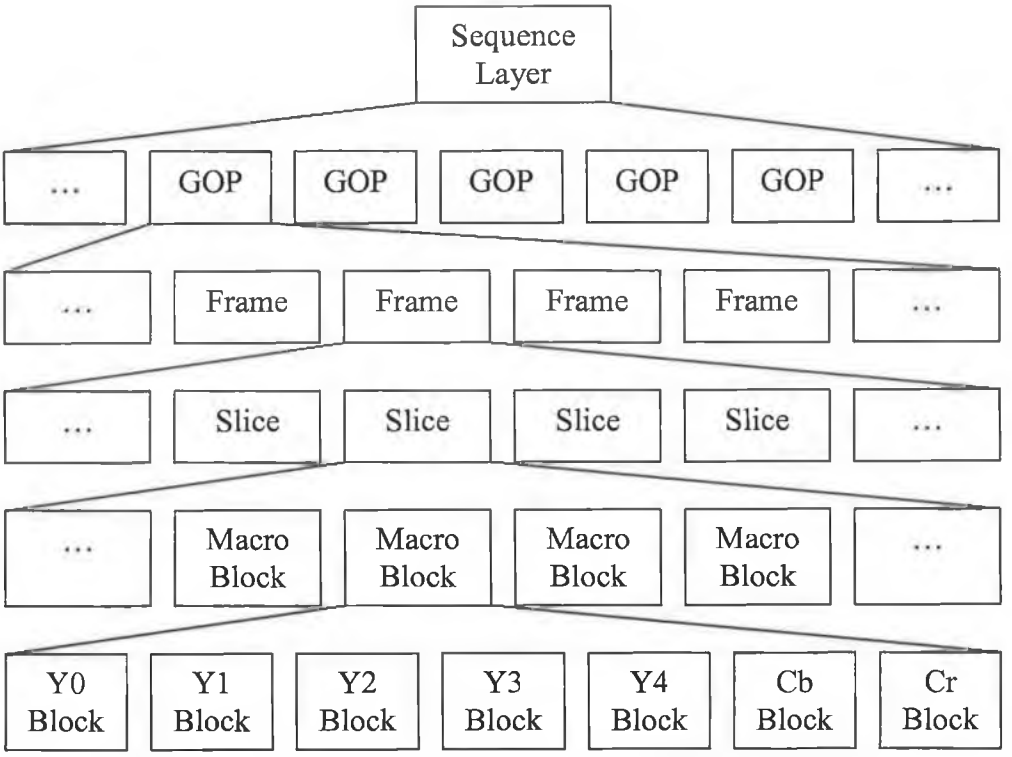


Figure 2-9. Layered structure of the MPEG bit stream.

Table 2-2. Function of each layer of the bit stream

Layer	Function
Sequence Layer	Contains general information about the video, such as vertical and horizontal size, aspect ratio, picture rate, buffer size and intra and non-intra quantizer default tables.
Group of pictures (GOP)	Pictures are grouped together for greater flexibility and efficiency.
Picture	Picture or frame layer is the primary coding unit that contains information regarding the picture's position in the display order, what type of picture it is (Intra, Predicted, or Bi-directionally), and the precision and range of motion vectors in the frame.
Slice	The slice layer is for the handling of errors.
Macroblock	Macroblock layer is the coding unit where motion vectors are stored.
Block	Block layer contains the coefficients of the pixels and it the smallest coding unit.

2.4.2 I, P, and B Frames

There are three types of frames as shown in Figure 2-10, these are determined according to the frame of reference they are compared (or predicted) to. Each frame is divided up into differently sized slices. Therefore a slice may contain one or all of the macroblocks in the frame.

- Intra frames (I-type) - An Intra or I frame is simply a frame coded in its complete still image form. This is used as a reference for the other types of frames.
- Predicted (P-type) - Predicted or P frames are frames encoded using a past (forward predicted) I or P frame as reference.
- Bi-directionally predicted (B-type) - Bi-directionally predicted or B frames are frames encoded using a past (forward predicted) and a future (backward predicted) I or P frame as a reference.

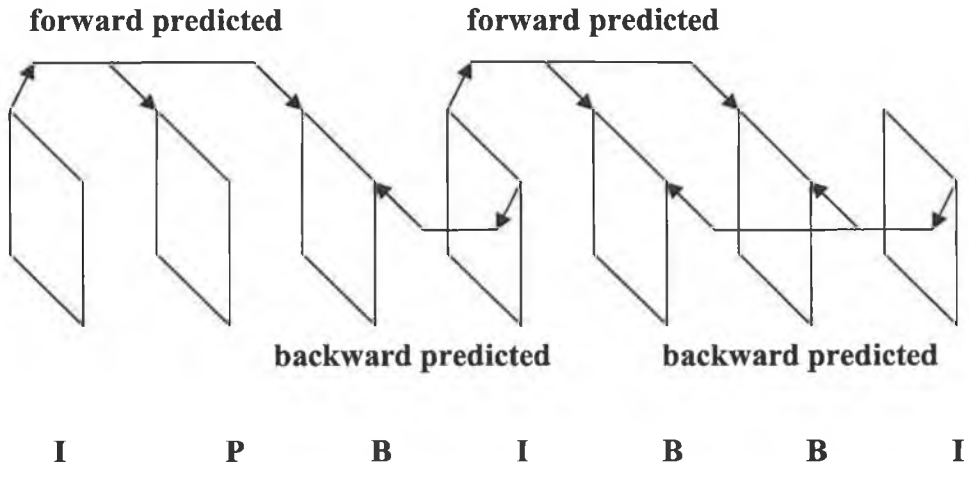


Figure 2-10. Illustration of I, P and B frames in forward and backward prediction.

However, MPEG doesn't store the sequence of frames in the bit stream in the exact order it is to be displayed, therefore the order of the frames needs to be decoded before it can be displayed. For example, Figure 2-10 gives a display order of:

I	**P**	**B**	**I**	**B**	**B**	**I**	**(frame)**
1	2	3	4	5	6	7	(index)

In order to decode B(3), I(4) needs to be decoded first. B(5) and B(6) also need I(7) to be decoded first before they can be decoded. Therefore the decoding order is as shown below:

I	**I**	**P**	**B**	**I**	**B**	**B**	**(frame)**
1	4	2	3	7	5	6	(index)

2.4.3 Macroblock

A macroblock is the basic unit in the MPEG stream. MPEG specifies a macroblock as an area of 16 by 16 pixels. MPEG also specifies that the data, which is represented in the RGB colour space (as discussed earlier), is to be converted to an YCbCr colour space. This colour space conversion is also the first step of MPEG compression as it down sample the stream through chroma subsampling. The resulting macroblock can thus be described with 4 luminance (Y) information, and 2 chroma (Cb, Cr) information (4:2:0). This subsampling does not affect quality as the human eye is less sensitive to chroma than luma information. A single macroblock is shown in Figure 2-11.

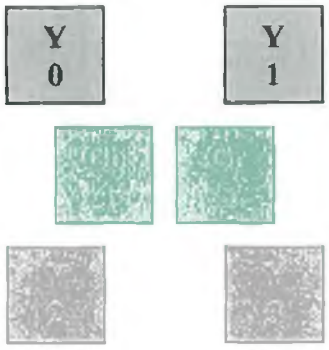


Figure 2-11. Structure of a macroblock with 6 blocks and its numbering convention.

Each macroblock is composed of six blocks made up of 8 by 8 pixels, which are transformed by Discrete Cosine Transform (DCT) to obtain the coefficients for the luminance and chrominance information. The DCT (a DFT related transform, Section 2.2.6. Fast Fourier Transform) transforms the values into the frequency domain in order to optimize the quantization performance that follows. The DCT coefficients after quantization will have a lot of zeros. As shown in Figure 2-12, a zig-zag scan is performed to re-arrange the values more efficiently for the next step in compression. For the first value termed the DC coefficient, a predictive coding scheme called Difference Pulse Coded Modulation (DPCM) is performed, while for the rest (AC coefficients), a Run-Length Encoding (RLE) scheme is performed. DPCM, a variant of PCM (refer to the previous section on Pulse-Code Modulation), encodes the values as differences between the current and previous value. RLE works by counting the number of similar values in a consecutive sequence in the series of DCT coefficients and replacing it with a count instead.

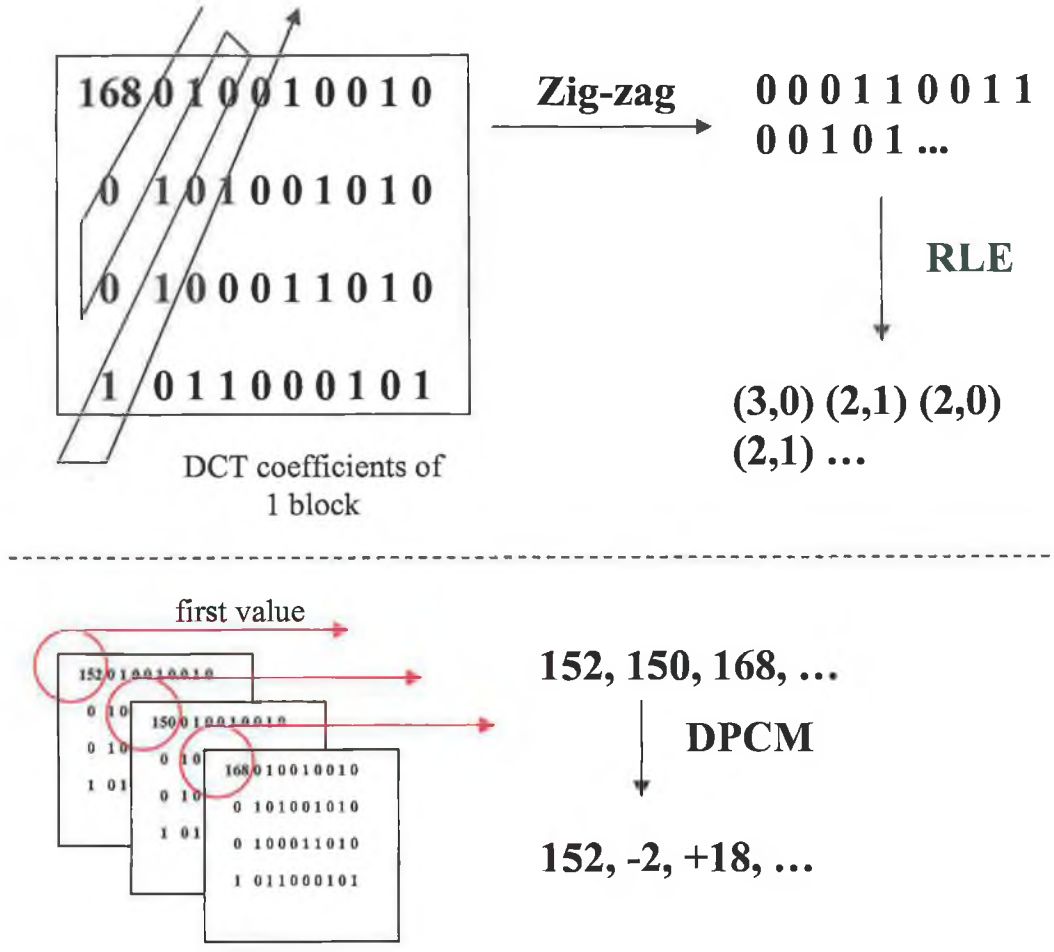


Figure 2-12. A zig-zag scan, DPCM and RLE is performed on the DCT coefficients of a block to output a compressed representation of the block.

2.4.4 Variable Length Coding

Depending on the type of frame being coded, there are different types of macroblock. Variable Length Coding (VLC) is used to encode the DC and AC coefficients to reduce further the spatial redundancy of each macroblock. VLC is a statistical coding technique that uses short codewords to represent values that occur frequently and long codewords to represent values that occur less frequently and is therefore used in almost every aspect of

MPEG-1. Thus encoding is done by referring to the VLC code tables as shown in Tables 2-3, 2-4, and 2-5.

Table 2-3. Macroblock types in an I frame.

Type	VLC code	MB quant
Intra-d	1	0
Intra-q	01	1

Table 2-4. Macroblock types in a P frame.

Type	VLC code	Intra	M F	Coded pattern	Quant
pred-mc	1		1	1	
pred-c	01			1	
pred-m	001		1		
intra-d	0001 1	1			
pred-mcq	0001 0		1	1	1
pred-cq	0000 1			1	1
intra-q	0000 01	1			1
Skipped					

Table 2-5. Macroblock types in a B frame.

Type	VLC	Intra	M F	M B	Coded pattern	Quant
pred-i	10		1	1		
pred-ic	11		1	1	1	
pred-b	010			1		
pred-bc	011			1	1	
pred-f	0010		1			
pred-fc	0011		1		1	
intra-d	0001 1	1				
pred-icq	0001 0		1	1	1	1
pred-fcq	0000 11		1		1	1
pred-bcq	0000 10			1	1	1
intra-q	0000 01	1				1
Skipped						

Key for the abbreviations used in Table 2-3, 2-4, and 2-5:

- VLC – variable length code
- M F – motion forward
- M B – motion backward
- pred – predictive
- m – motion compensated

- c – at least one block in the macroblock is coded and transmitted
- d – default quantizer is used
- i – interpolated, a combination of forward and backward prediction
- b – backward prediction
- f – forward prediction
- Skipped – the macroblock is the same as in the previous frame

2.4.5 Motion Estimation and Compensation

Motion estimation and compensation are the underlying techniques that allow MPEG-1 to achieve very high compression rates. The difference between successive frames is small, therefore the difference between them is encoded. The method to find these differences and using them to reconstruct the video is called motion estimation and compensation.

The aim of motion estimation is to search for the macroblock in the next (search) frame that matches the macroblock in the reference frame. I or P type frames can be used as the reference search frame for either P or B type frames. Forward motion estimation is achieved when a previous reference frame is used to predict a future frame, while backward motion estimation is achieved when a future reference frame is used to predict a previous frame.

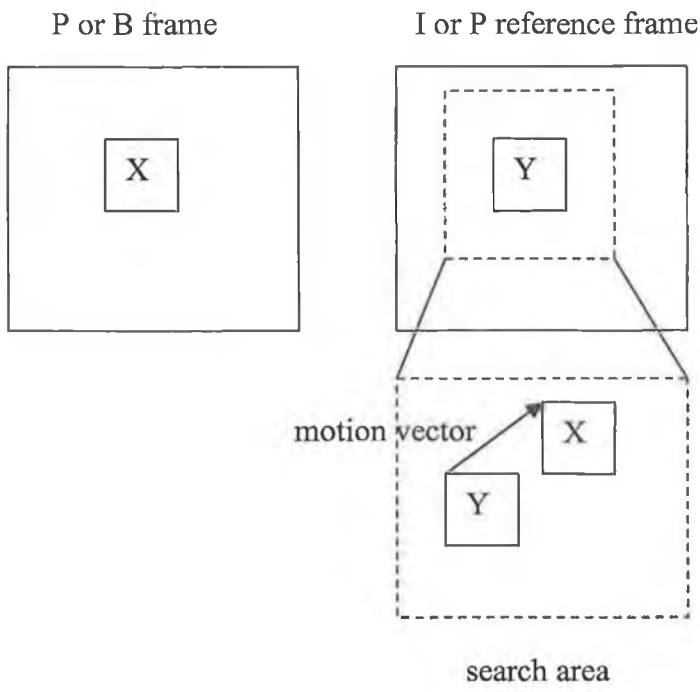


Figure 2-13. A motion vector is calculated by searching in a reference frame for a Y macroblock that matches the X macroblock.

Figure 2-13 shows how a motion vector is derived from the reference frame by calculating the difference between macroblock X and Y. There are many block matching algorithms that can be used to search and estimate the motion vectors, each with their trade-offs in accuracy and processing time. A full search strategy where every macroblock is compared with every possible macroblock, although providing the best match, takes too much processing time. However there are numerous algorithms such as three-step search, algorithmic search, and diamond search [Barjatya, 2004] which significantly reduce the computation time at the cost of a small reduction in accuracy. The savings in computation time are mainly made in cutting out unnecessary computation by only performing more detailed searches at regions where the probability of finding a best match is at its highest. The various search strategies such as three-step, algorithmic, and diamond search are based on this same idea with different implementations that yield different results. MPEG-1 does not specify any particular algorithms, therefore it is up to developers of MPEG encoders to use their own algorithms.

2.5 MPEG-1 Audio

2.5.1 Introduction

The audio encoding part of MPEG-1 is defined in the MPEG-1 standard ISO/IEC 11172-3. MPEG audio compression exploits the psychoacoustic model of the human auditory system. Much of the compression removes parts of the audio signal that are perceptually irrelevant to human listeners therefore achieving high lossy compression rates without much degradation in perceived quality. The following are some of the features of MPEG audio:

- Sampling rate – the sampling rates can be 32 KHz, 44.1 KHz, or 48 KHz.
- Audio channel support – MPEG audio can support one or two audio channels in four possible modes
    - Monophonic mode for a single audio channel
    - Dual-monophonic mode for two independent audio channels
    - Stereo mode for stereo channels that share bits
    - Joint-stereo mode that takes advantage of correlations between the two channels or the irrelevancy of the phase difference between the two channels.
- Predefined bit rates – the bit stream can be predefined into fixed bit rates ranging from 32 Kbps to 224 Kbps per channel.

2.5.2 Layer I, II and III

MPEG audio offers a choice of 3 independent layers of compression. This is intended to provide a wide range of trade-offs between complexity and audio quality. Layer I is the simplest layer and is suited to bit rates above 128 Kbps per channel, and is suited to high quality audio applications. Layer II introduces more complexity and is targeted for bit rates around 128 Kbps per channel applications. Layer III, which is also the most popular

(MP3), is the most complex scheme but offers the best audio quality targeted at around 64 Kbps per channel applications.

2.5.3 Auditory Masking

The human auditory system is unable to hear weak audio signals that are masked by strong audio signal called auditory masking. This masking effect makes weaker audio signals in the temporal or spectral neighbourhood imperceptible in the presence of a strong audio signal.

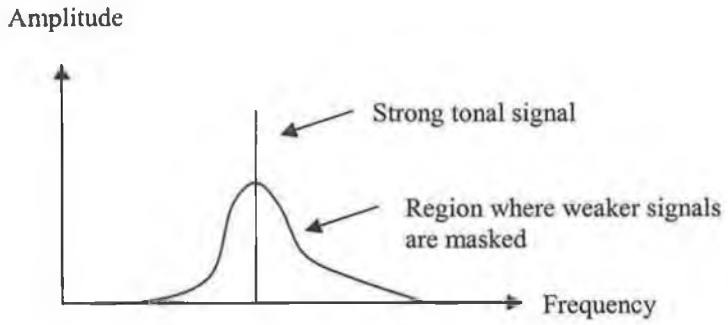


Figure 2-14. Audio noise masking.

Figure 2-14 shows the region where weaker signals are masked in the presence of a strong signal. The bulk of MPEG audio compression is achieved by removing the weaker parts of the audio signal.

2.5.4 MPEG Audio Encoder and Decoder

Figure 2-15 below illustrates an MPEG encoder and decoder [Pan, 1995].

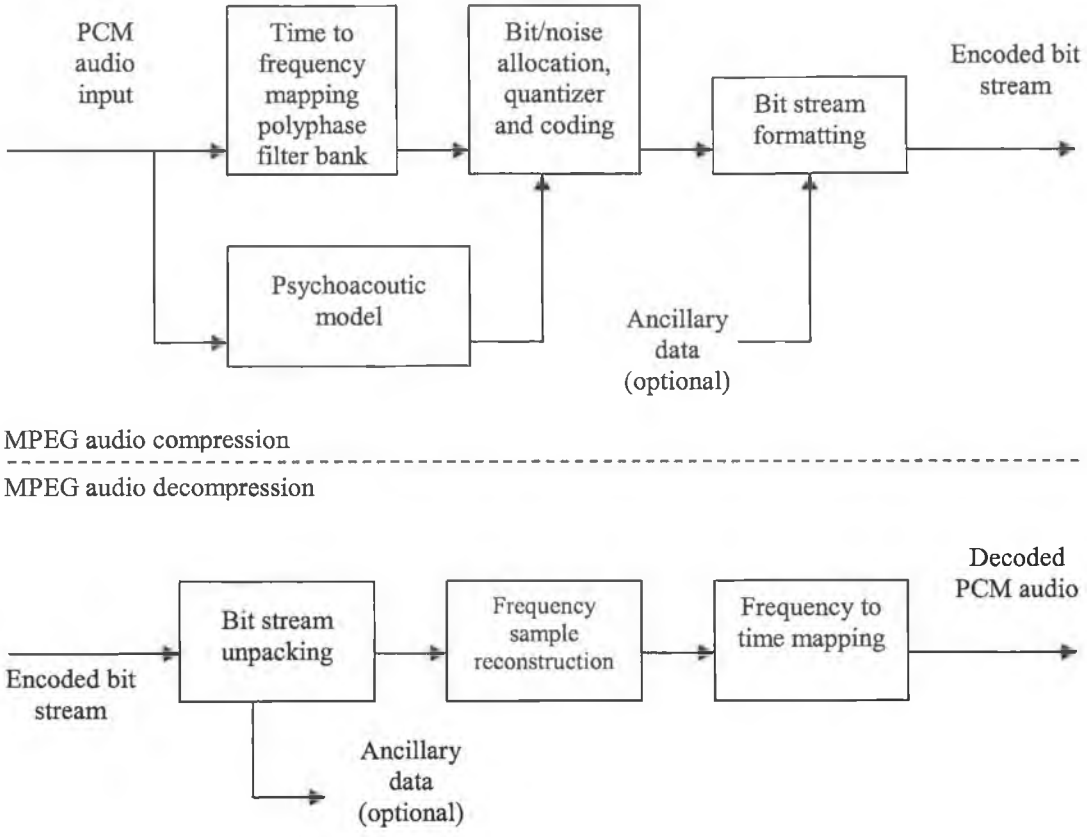


Figure 2-15. MPEG audio compression and decompression.

The signal in the form of uncompressed PCM audio input passes through a polyphase filter bank that divides it into 32 equally-spaced subbands of frequency. At the same time, the signal is also passed through a psychoacoustic model to determine the ratio of the signal to the masking threshold for weaker signals for each subband. This is intended to model the critical bands of human hearing where it is at the most sensitive.

The psychoacoustic model block calculates the signal to mask ratio and passes it to the next block for quantization. The audio signal is passed through an FFT function to derive a high quality frequency domain representation for each subband. From there, tonal and non-tonal components of the audio are identified because the masking effects between the two are different. The masking threshold is then established for each subband and used to calculate the signal to mask ratio.

The bit or noise allocation block uses the signal to mask ratios from the psychoacoustic model block to decide how to apportion the total number of code bits available for quantizing the subband signals. Depending on the layer, the quantization algorithm used is different. Layer I codes audio in frames of 384 audio samples by grouping 12 samples from each of the 32 subbands. Each group of 12 samples gets a bit allocation and if it is zero, a scale factor. This scale factor is a multiplier that sizes the samples to fully use the range of the quantizer. This quantization in MPEG audio Layer I provides a very large dynamic range of more than 120 decibels (dB). Layer II enhances the Layer I algorithm by coding data in larger groups with some restrictions on the possible bit allocations for the middle to higher subbands, along with a more compact code for the bit allocation scale factors. Layer III is also based on the Layer I and II algorithms, but further refines them by compensating for the filter bank deficiencies by processing the filter outputs with a Modified Discrete Cosine Transform (MDCT).

The bit stream formatting block takes the quantized samples and formats them into a coded bit stream. For decompression, the coded bit stream is decoded by restoring the quantized subband values in order to reconstruct the audio signal.

2.6 Summary

In summary, this chapter introduced common digital video terms and concepts used in content-based video retrieval systems. Digital video is made up of a series of still pictures that are synchronized to audio. The method of browsing digital video is typically based on the physical tape video player metaphor, which does not take advantage of the random access function of digital video. Digital video can be converted between different colour spaces that can be used to extract colour information from the video to be used in an affect-based indexing and retrieval system. Digital video is typically compressed using a compression format such as MPEG to reduce its file size.

Chapter 3 Content Based Video Retrieval

The discussion of content-based video retrieval (CBVR) systems in this chapter is developed within the context of affect-based indexing and retrieval as presented in this thesis. The main goal of affect-based content management is to allow users to search for videos based on its detected features. Therefore it is useful to understand the purpose and challenges facing existing modern CBVR systems.

3.1 Introduction

The role of a CBVR system is to process and store large amounts of digital video files for easy access in response to a user's information need. The system automatically generates metadata for videos based on a number of content analysis methods. These metadata holding information describing the video are the key component in locating the desired piece of video from within large archives.

As shown in Figure 3-1, users interact with a user interface component of the system to express their video information need and for the results of the retrieval process to be displayed. Users express their needs by entering a text or visual query example that matches the description of the video they are looking for. The system then takes this query and compares it to the video metadata and returns a list of best matches for the user to browse.

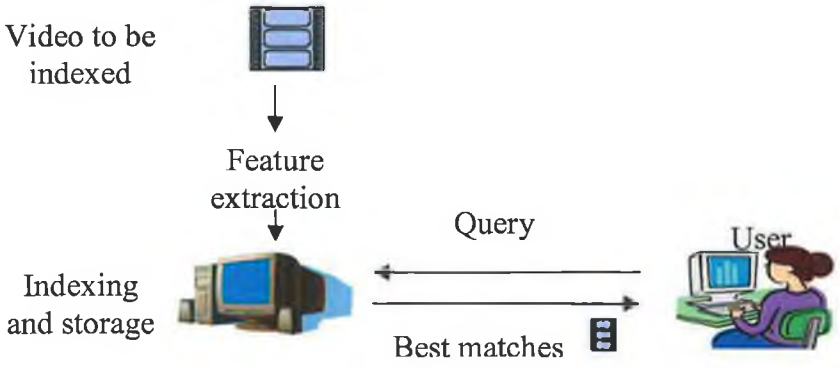


Figure 3-1. An overview of a typical video retrieval system.

Much work into multimedia information retrieval (IR) was based on image retrieval systems which focused on feature-based similarity search over images. Detected low-level features from the images such as colour, texture, and shape were used to annotate the images. These are stored as metadata which are then matched in terms of similarity against a user's text or visual input query. Image retrieval systems such as IBM's query by image content system (QBIC) [Flickner et al., 1995] and Virage image search engine [Bach et al., 1996] have gone on to influence modern video retrieval system's method of searching and browsing video material as observed by Lee [1999].

3.2 Challenges for Content-Based Video Retrieval Systems

In the broader context of creating a CBVR system, there are numerous challenges, which are reflected in large video retrieval evaluations (TRECVID) [Over et al., 2006]. TRECVID originates as a track of the Text Retrieval Conference (TREC) but was separated into a separate workshop in 2003. The aim of TRECVID is to provide common large test sets and uniform evaluation procedures for exploration and testing of technologies for CBVR, as well as a forum for disseminating participant's results [TRECVID]. The challenges of CBVR can be broken down into a few categories.

- **Feature Extraction**

Choosing the correct set of features for processing is one of the challenges facing any CBVR system designer. The selection of relevant and useful features is important as it directly affects the reliability and effectiveness of a CBVR system. Ideal features would be those that are highly discriminant in terms in filtering out unnecessary information or noise. Special considerations regarding processing time also have to be factored in for implementation purposes. Low-level features are combined or processed to form higher level features to aid tasks such as shot boundary detection and detecting camera motion. Some of the common low-level features used in CBVR systems are motion vectors, texture, colour histograms, and shapes. However, low-level features are semantically weak, making it difficult for users to perform retrieval at a higher level of abstraction. As

one feature may perform well for some specific tasks, in reality a combination of features is usually used to achieve a higher level of semantic abstraction.

In order to fulfill the needs of affect-based video indexing and retrieval, this thesis presents a review and explanation of techniques for feature extraction of affective labels in Section 6.4 (Feature Extraction).

- **Video Retrieval Engines for Very Large Video Collections**

Text-based web search engines such as Google, Yahoo and MSN search receive millions of queries from users every day. In order to satisfy their users they must present accurate results in at most a few seconds. Their web crawlers have to store, index and retrieve billions of documents every day. If video retrieval systems are to be deployed at such a scale, there will be challenges to overcome such as storage, bandwidth, feature extraction, indexing, and search times. However, these are implementational issues which can be solved by distributing the processing across a vast number of machines.

- **Simple and Intuitive Interface for Users**

Searching and browsing videos can be a tiring task. It is essential that the interface used in CBVR by users does not distract the user's concentration. The system has to present results in a clear format that is easy to understand. The popularity of web-based search engines suggests that simplicity of use is very important. Users typically do not use many search terms and only browse through at most the first few pages of results [Silverstein et al., 1999]. It is essential to make the query and results as easy to understand as possible, especially when non-expert users are involved. Technical terms such as "shot cut" or "keyframe" are often used in a CBVR system, therefore there is a need to translate those terms into easier to understand words. This critical aspect concerning CBVR systems and users is identified as the "semantic gap", a well-known problem image and video retrieval [Smeulders et al., 2000], which is discussed further in Section 3.3 (Bridging the Semantic Gap with Affective Computing).

This thesis investigates the application of affective computing techniques to facilitate user search of videos by offering them a new interaction possibility using emotional labels inferred from affective labels. This system is presented in Section 6.7 (Verbal Labeling).

3.3 Bridging the Semantic Gap with Affective Computing

Lee et al. [1999] observed that while early video retrieval systems were based on image retrieval systems, they were not as intuitive or user-friendly as expected especially for users who are non-experts in IR. The reliance on feature-based similarity search in early systems meant that users had to have some understanding of the semantics surrounding the low-level features that were being used. The terms describing low-level computational features might be too technical or unsuitable for these non-expert users to use effectively.

As a crude example, non-technical users might understand and be able to generate higher-level queries such as "exciting car chase scenes" more easily than a query such as "rapid shot cuts + large motion activity + presence of car object" that an expert user could generate to find an action scene from a movie; thus the need to bridge the "semantic gap" together.

Lew et al. [2006] described the problem as "systems built by research scientists were essentially systems which could only be used effectively by scientists". Their study into the state-of-the-art and challenges in multimedia IR discusses the point that for video retrieval systems to be useful to casual users, a human-centered approach should influence the design of modern retrieval systems.

They highlighted the emerging field of affective computing as one of the approaches that holds promise for human-centered systems. Affective computing seeks to understand user's emotional state and to respond accordingly as an intelligent and context-aware system. For example, using an affect-based video retrieval system users might find that formulating queries such as "exciting scenes" is much easier and more helpful than trying to define it in terms of "rapid shot cuts and elevated audio energy".

This higher level of abstraction allows users more flexibility to concentrate on reviewing the returned results.

3.4 The Three Levels of Content Retrieval

As modern CBVR systems are based on earlier image retrieval systems, some of the lessons learnt in image retrieval can be applied into video retrieval as well. There are many ways to perform retrieval based on the many attributes that can be detected from video. [Eakins, 1996] in his review of automatic image content retrieval reasoned that these can be divided into 3 levels. The three levels are:

- Level 1: the lowest level comprises of primitive features such as colour, texture, shape, spatial location or a combination of the above. Retrieval is at its most basic level where an example of a query would be "find blue and yellow regions" or "find a texture similar to that of a woven cloth".
- Level 2: comprises of derived attributes or "logical" features that involve some degree of logical inference about the identity of objects. This is a higher level of abstraction, but the retrieval queries are still clearly defined semantic units such as "find red car" or "find faces of people".
- Level 3: comprises of abstract attributes that involve a high degree of abstraction and possibly subjective reasoning about the meaning or purpose of the objects or scenes depicted. Some examples of queries at this level are "find a calm scene" or "find a funny looking face". At this level, subjective reasoning is important, therefore the accuracy or relevancy of a retrieval task result can be subjected to individual interpretation. While results could vary from one person to the other, nonetheless, when faced with a large volume of data, retrieval at such a high level of abstraction can still be useful in filtering out a lot of unnecessary data.

These 3 levels in a video retrieval context are also observed in Hanjalic and Xu [2003] in which they are described as the 3 levels of video content perception, analysis and retrieval. Level 1 could then be described as the Feature Level, Level 2 as the

Cognitive Level, and Level 3 as the "Affective Level". As can be seen in Figure 3-2, the higher the level, the more abstract and subjective the information need is.

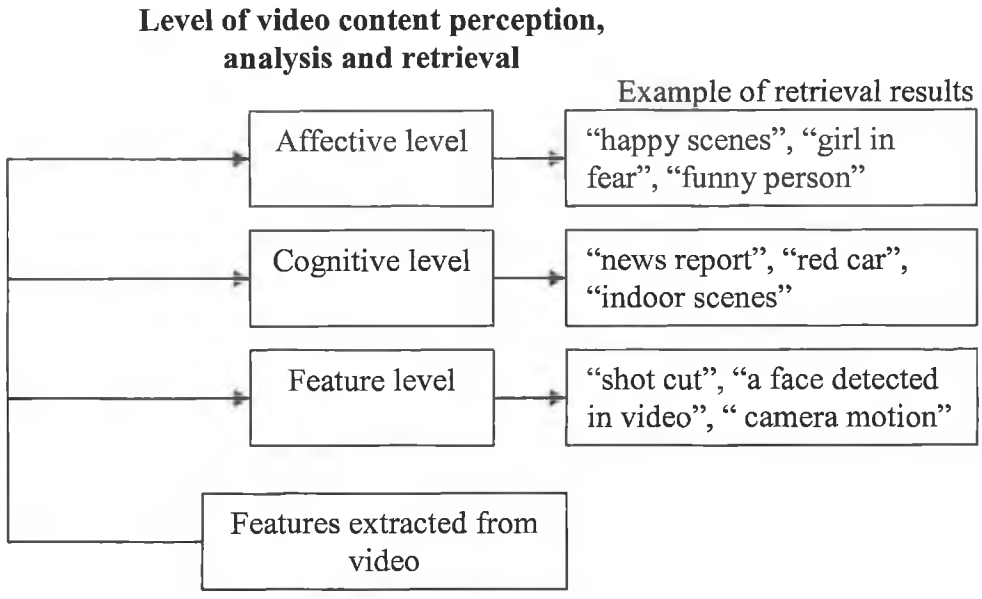


Figure 3-2. The three levels of video content perception, analysis and retrieval by [Hanjalic and Xu, 2003].

3.5 Examples of Video Retrieval Systems

Modern examples of video retrieval systems try to address the problem of the semantic gap by exploring different strategies of personalization, visualization, and summarization [CDVP] [Informedia]. Before exploring the background and issues of Affective Computing in more detail, the remainder of this chapter reviews some significant examples of previous work on content-based multimedia information retrieval. These video retrieval systems are chosen as exemplars of the many examples developed around the world [Smeaton, 2004].

In CBVR, video data such as TV news and entertainment segments is typically used as the data set. TV represents a compelling source of video data for CBVR systems since it is easy to capture using a desktop computer equipped with a TV tuner device and applications can be envisaged for home users. In addition, the rich amount of aggregated

information presented in the news segments in a rigid format makes the information contained in it useful for research and exploration of CBVR technologies. Feature detectors in a CBVR system typically take advantage of this information to detect different scenes, the anchorperson or on-screen text and graphics. Even in entertainment shows such as sitcoms and documentaries there are general trends or structure that can be captured or inferred and used in the feature detectors of a CBVR system. The retrieval systems described in the following subsections typically use features such as speech transcriptions to allow users to search and retrieve videos using text queries.

In recent years there has been growing interest into other domains of video data such as sports video and films. The potential value of being able to do content-based retrieval on these domains of video data is huge as sports and films are two of the most popular forms of entertainment medium.

- **Medusa Multimedia Retrieval System**

One of the earlier video retrieval systems was developed by Cambridge University and Olivetti Research Limited is the Medusa multimedia retrieval system [Brown et al., 1995]. This was based on combining related work on the Video Mail Retrieval (VMR) and Medusa project, which were also developed by Cambridge University and Olivetti Research Limited. The Medusa multimedia retrieval system project was designed for the automatic content-based retrieval of broadcast news.

The Medusa project provided the network infrastructure necessary for sending many streams of digital video and audio over a prototype 100 Megabit-per-second switched Asynchronous Transfer Mode (ATM) network. The video retrieval techniques were based on the VMR project which developed an application that was able to transcribe and store the uttered words in video mail messages using speech recognition and store them in a searchable system that allowed fast and easy retrieval utilizing text retrieval techniques [Brown et al., 1994]. The Medusa television news broadcasts retrieval system stored captured video and audio data along with teletext information that contained transcriptions of the words spoken. User's retrieval needs were then entered using a typed text request (query). Given the textual query, the statistical text retrieval

technique Okapi BM25 model was used to generate a list of results that presented the closest possible match to the user's query. The BM25 text retrieval model was chosen as it had been shown to be effective on extremely large text corpora [Sparck Jones, 1994].

- **The Físchlár Digital Video Library System**

The Físchlár digital video library system started development in 1997 in the Centre for Digital Video Processing (CDVP) at Dublin City University to showcase its research work. It was designed as a complete end-to-end system that supports capture, indexing, browsing, and searching through archives of digital video information and has been deployed onto four separate digital video libraries [Smeaton et al., 2004]. The four digital video libraries collections are comprised of TV programs, TV news programs, the TREC video tracks, and educational nursing videos.

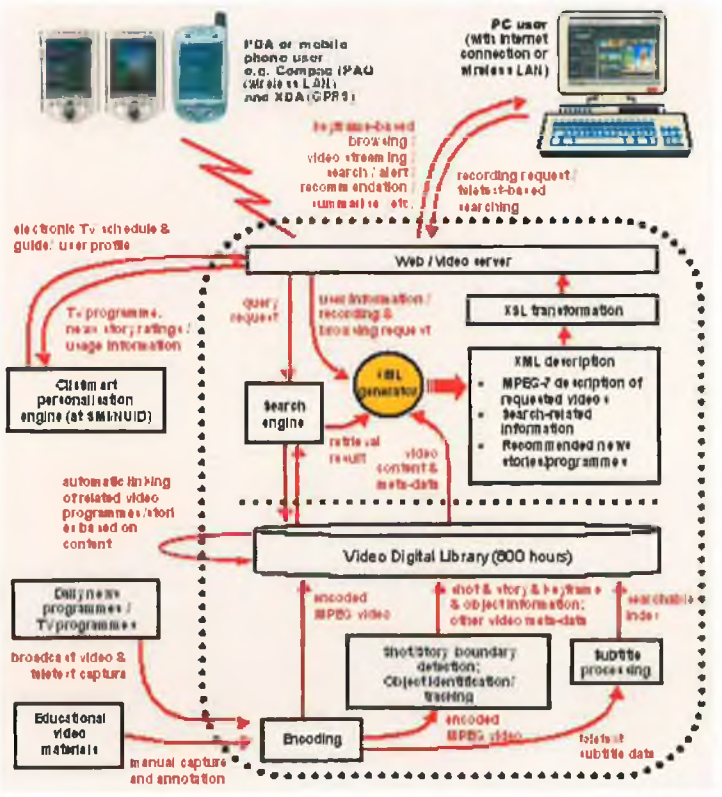


Figure 3-3. Architecture of the Físchlár digital video library system, taken from [CDVP].

As shown in Figure 3-3, Físchlár performs segmentation of videos (in MPEG-1 format) by detecting shot boundaries and selecting the representative keyframes from these shots [O'Toole et al., 1999]. A transcript of the sentences being spoken is detected either directly through the teletext (closed-captions for TV-captured video) or through the use of speech recognition software. Feature analysis is also done by extracting features such as speech music discrimination [Jarina et al., 2001] and recognition of on-screen faces [Czirjek et al., 2003]. Users of Físchlár browse the videos by searching it using text queries augmented with keyframe-browsing and video summarization and recommendation features.

Figure 3-4. Físchlár TV video browser, taken from [CDVP].

Físchlár facilitates a human-centered search environment through personalization of the user's information request. The user interface as shown in Figure 3-4 plays a large part in simplifying searching by streamlining the visualization and search task procedures. Users can type in a text query and optionally click on radio buttons to include features such as speech or music for selection. Small icons beside the returned results representing the various features give the user a quick overview of the video content. Físchlár uses an underlying architecture which stores derived video metadata in MPEG-7 and supports Extensible Markup Language (XML) based processing of the information. The use of XML allows the video content information to be processed into standard Hypertext Markup Language (HTML) that is readable by web browsers. The video itself

is stored on streaming video servers and is fetched when playback is needed as users browse through the video collection. The HTML-formatted information enables Físchlár users to access personalized video information through web browsers and web-enabled portable devices such as personal digital assistants (PDA) [Lee et al., 2000].

- **Informedia Digital Library System**

The Informedia Digital Library system is another well-known system developed at Carnegie Mellon University (CMU) starting in 1994. The Informedia system as shown in Figure 3-5 pioneered video retrieval techniques by combining video with audio and other features on a large archive of broadcast TV news and documentary content. The system's success was much attributed to the use of textual information through speech analysis that can be derived from video for search and retrieval.

Figure 3-5. A text query and result user interface from Informedia, taken from [Informedia].

[Hauptmann, 2005] explained that speech analysis using a highly accurate speaker-independent speech recognizer combined with closed caption text can produce a transcription that is meaningful and highly relevant to the events in the video, especially in the domain of broadcast TV news. From the visual domain, face detection and video optical character recognition (OCR) provided additional textual data.

Well-established text indexing techniques are used on the transcribed speech to index the video. This has been proven to be useful in further studies that reveal non-expert users do prefer systems that have textual query capability such as done by [Christel and Conescu, 2005]. The other significant factor in the success of the Informedia system is attributed to the quality of their user interface. Much effort has been put into video information summarization and visualization, starting with shot boundary detection technique based on colour histograms for keyframe selection. A strategy of video "skimming" is then employed to characterize scenes using a combination of visual, audio and object features such as audio level, audio transcripts, camera motion, face detection and text detection [Smith and Kanade, 1995]. This creates a compact storyboard that allows user to quickly browse or find what they were looking for.

3.6 Summary

In summary, this chapter introduced CBVR systems and how it works. A typical CBVR system automatically indexes video with metadata and stores them in a database. When a user enters a text or visual query, it is compared in terms of similarity to the metadata that were stored in the database. The system would then return a video that has the best match between the metadata and the query. A review of some of the challenges and issues facing modern CBVR systems that has to be considered for the development of an affect-based retrieval system is also presented. Many attributes can be detected from video, these can be divided into 3 levels. The highest level (Level 3 or the Affective Level) is the most abstract and subjective level that is currently being addressed in modern CBVR systems in order to bridge a semantic gap between system-detected features and user's queries. The chapter concluded with a few examples and explanations of modern CBVR systems.

Chapter 4 Affective Computing

The affect-based video indexing and retrieval system presented in this thesis is largely inspired by the emerging field of affective computing. Recent interest and developments in affective computing have allowed it to be applied to CBVR. This chapter examines current and emerging work in the field of affective computing and explains why it is relevant in a CBVR context.

4.1 Introduction

Affective computing is computing that "relates to, arises from, or deliberately influences emotions" [Picard, 1997]. The word affect is a measure of how a stimulus could have an emotional or cognitive impact upon humans. Affective computing is a relatively new field of computing fraught with many new challenges. The central tenet of the field is the inclusion of emotions into the computing process that involves human users. In her pioneering book on affective computing Picard [1997] argues that emotions, though difficult to characterize and study, cannot be ignored in the field of computing as they play an important part in human perception, communication and decision-making.

Due to the inherent difficulty of understanding and directly measuring the emotions that humans are experiencing, much work into affective computing is focused on modeling and detecting affective features from physiological or video and audio data. This can be thought of as measuring what are the key stimuli on a person that are trying to elicit a feeling or emotion, as opposed to measuring what the person is currently feeling. This is necessary and more practical because it is difficult to determine the variation in how each person might interpret or perceive an affective event. For example, a simple joke could be funny or not funny according to whom the joke is told to. Instead of trying to measure and anticipate how each person will react to the joke, it is far more practical and consistent to measure what the joke is trying to do (make a person laugh or subtle criticism).

Although there are also variation in how affect can be expressed (such as the way a joke is being told that might make it funny or not), research in affective computing shows that these problems can be overcome. For example, affective features can come from a variety of different sources such as human body language, facial expression, and speech intonation. On its own, each affective feature might not be enough to detect the underlying emotional state, however these features could be combines to understand and estimate or predict the underlying emotional state. This issue is addressed in more detail in Section 4.4 (Addressing the Criticisms of Affective Computing).

Therefore, features in data that are shown to be key influences in affect are called affective features. Affective features can also be described as the pieces of encoded information that can be said to elicit or be captured from an emotional response.

4.1.1 Why are Emotions Important?

At first, the prospect of including emotions into a machine's logic might seem to be counter-intuitive as emotions can be perceived to lead to irrational decisions. Many science-fiction writers have expressed doomsday scenarios when machines are intelligent enough to be able to perceive and feel negative emotions. However, the focus here is not to imbue machines with emotions but rather the tools for recognition and understanding of human emotions. This ability allows a computer to analyze and react appropriately based on the emotional state of its human user.

- **Human Perception**

Drawing from neurological studies Picard [1997] inferred that emotion greatly influences human perception. Over the course of Cytowic's [1989] study into the synesthetic experience, it is demonstrated that the limbic system of the brain, which is the seat of emotion, memory and attention is highly activated during sensory perception. This shattered the previous assumption that the limbic system plays a lesser role compared to the cortex, which is regarded as the home of sensory perception. Izard [1993] also

concluded in his treatise on emotion theory that emotion is both a motivating and guiding force in human perception and attention.

- **Human Decision-Making**

Human decision-making is often stereotyped as falling between the "thinking" and "feeling" category. Therefore humans are said to make impulsive decisions based on feelings and rational decisions based on rational thought. However, neurologically, there is no evidence of a clean dividing line that separates the two. All sensory inputs in the human brain must pass through the limbic part (emotion, memory, attention) before being redistributed to the cortex (sensory perception) for analysis. This suggests that the analysis of the senses could lead to a decision that is very much connected to emotions.

On the other hand, the lack of emotions could actually impair the process of decision-making. Damasio's [1997] study into patients with frontal-lobe disorders found that people who suffer such disorders have difficulty making decisions such as choosing a date for an appointment or meeting. Patients who suffer from frontal-lobe disorders that affect the limbic system governing emotions are found to take more time to make such decisions, if they can make any decision at all. Damasio hypothesizes that emotions play a biasing role in decision-making, establishing the values used in evaluating potential outcomes to ward off an infinite logical search. These findings provide neurological support that emotions are vital for healthy and rational human thinking.

- **Human Affective Communication**

Emotions are a natural and social part of human to human communication. We use subtle body language and facial expressions to help us communicate better and more efficiently. Slight nods of the head when conversing with someone can let them know when to proceed or can also be used as a sign of acknowledgement. Increasingly, more people are interacting with the computer, sometimes more than with humans. Through communications technology, we now use e-mail, instant text messaging clients and

internet voice calls to keep in touch with other people. Even behind a computer terminal, humans still use emotions to communicate effectively.

A study by [Nass et al., 1994] established that human's interaction with computers is as natural and social as with humans. They conducted the study by performing a number of classical studies of human social interactions, and then substituting computers into a role usually occupied by humans. This was meant to compare how human-human and human-computer interactions differ. For example, an experiment was set-up where person A had to rate person B's performance in tutoring or teaching, either directly face to face or through a third person. The results, as expected, were that people were nicer in person B's evaluation when it is conducted face to face. But when person B is replaced by a computer, it is found that the people were still being as nice as they were towards a human. The results of this study show that humans still use emotions even when interacting with computers. This result might not be too surprising as there are many instances of people expressing anger at their computer when it breaks down.

The popularity of people using "emoticons" or "smileys" in informal e-mails and text messaging also shows how affective cues can be used to augment and improve on the weaknesses of standard textual communication. Emoticons or smileys are text or graphical shorthand, Table 4-1 shows examples of emoticons indicating specific emotion or action being simulated by the writer. These emoticons give the reader a glimpse into the writer's emotional state or intention which might be difficult to convey in written words. Misunderstandings can be averted with just a simple "smiley" face to emphasize that a statement is a joke and not to be taken too seriously.

Table 4-1. Examples of "emoticons" or "smileys" and their meanings.

Emoticons or Smileys	Meaning
☺	A smiling face
☹	A sad face
:-]	A polite smile
;-)	A wink
:-O	A shocked face
>:-O	An angry face

4.1.2 Research into Human Emotion

Emotion is complex and there is no universally defined definition for it. Research into human perception of emotions and its underlying processes is an extremely complex and wide field. Therefore leaving aside too much detail into defining emotions, we are interested here in useable knowledge about them that can be used to implement a real system taking into account affect-related information. Much like how psychoacoustics is being used in audio compression such as MPEG, research into human physical and cognitive aspects of emotion can help us to model and detect affective features.

Cornelius [1996] in his review of contemporary theory and research on emotion psychology stated that there are four main theoretical perspectives. The first two perspectives (Cognitive and Darwinian perspective) form the bulk of the foundations of emotion psychology for affective computing. The other two are less important in affective computing. The four perspectives are explained in the subsections below:

- **Cognitive Perspective**

The Cognitive perspective makes the assumption that thought and emotion are inseparable, therefore emotions are dependent on appraisal, the process by which events in the environment are judged as good or bad. The brain undergoes a cognitive process of appraisal and the result gives rise to emotions. Therefore the cognitive process of thought, such as becoming aware of being in a dangerous situation, gives rise to fear.

Using this paradigm, a dimensional approach can be used to describe emotions by dividing it into a set of interrelated primitive appraisal components. The valence, arousal, dominance (VAD) appraisal components, proposed in [Osgood et al., 1957] and [Russell and Mehrabian, 1977] can be shown to be sufficient to characterize all emotions. The idea that emotions can be divided into 3 components for independent appraisal therefore is useful in affective computing because detected affective features can be characterized and mapped into these 3 components. The VAD appraisal components are explored in more detail in Section 4.2 (VAD Emotion Space).

- **Darwinian Perspective**

The Darwinian perspective introduced by Charles Darwin [Darwin, 1872] postulates that basic emotions are evolved phenomena via natural selection that allow important survival functions. It is thought that basic emotions such as fear were developed to help humans to survive by acting as a safety mechanism. This also means that emotions are biological in origin and are therefore universal among humans. Most research in this area is focused on physical displays of emotions including body language and facial expressions. The work of Ekman and his co-authors in determining universal and cultural differences by analyzing human facial expressions identified 6 basic emotions that were universal: Happy, Surprise, Anger, Sad, Fear and Disgust [Ekman et al., 1987].

- **Jamesian Perspective**

The Jamesian perspective proposed by William James in the 1800's views emotions from the perspective that it would be impossible to have emotions without bodily changes [James, 1890]. This meant that any biological or chemistry changes of the human body correspond with what emotions are being experienced. This can be demonstrated by the lie-detector test where changes in physiological signals might indicate a change in emotion.

- **Social Constructivist Perspective**

The Social constructivist perspective takes the view that emotions are cultural products that owe their meaning and coherence to learned social rules. Therefore emotions being experienced vary a lot depending on the culture and environment that a person experiences [Derry, 1999] [McMahon, 1997].

The cognitive perspective makes it easier to define emotions therefore the affect-based retrieval system presented in this thesis is focused on the cognitive perspective and

uses the VAD appraisal components to map system-detected affective features to emotional verbal labels.

4.2 VAD (Valence-Arousal-Dominance) Emotion Space

The VAD emotion space is a vital part of affective computing due to the way it can describe and represent emotions by breaking them down into 3 easier to define components. Osgood et al. [1957] and Russell and Mehrabian [1977] both suggested that 3 independent and bipolar dimensions can describe all emotions experienced by humans. Russell and Mehrabian's study involved a large number of human subjects being given a list of 151 emotion term or keywords that describe a distinct emotional state and asking them to rate each of them on a scale based on how they feel towards the emotion terms according to the 3 dimensions. The subjects' scores were then averaged to give valence, arousal, and dominance values for each of the 151 words. The three components are explained in the following subsections:

- **Valence**

Valence is a measure describing the level of pleasure to displeasure ranging from a "positive" response associated with extreme happiness or ecstasy through to a "negative" response resulting from extreme pain or unhappiness. This is shown in Figure 4-1, where the y-axis represents valence values, and the x-axis represents time. This plot of valence, which could be detected from a video, shows the development of valence over a period of time. When the valence values are high and sustained for a period of time, the video is said to contain high valence. For the affect-based retrieval system, the neutral feeling or "zero" value is derived from the average of the valence values by taking the assumption that long pieces of video, such as films, are said to contain equal amounts of positive and negative emotions due to their broad expressions of emotion. However, the limitation of this method is that not all pieces of video have equal distribution of emotions, leading to bias. Therefore the use of long pieces of video is aimed at reducing such bias. This

limitation should be explored as future work to reduce bias as well as to allow short pieces of video to be analyzed.

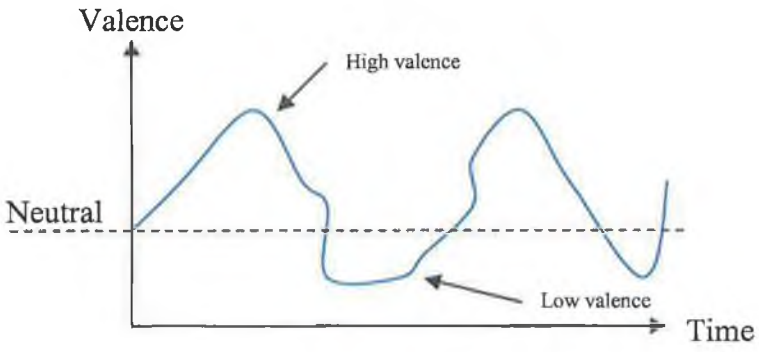


Figure 4-1. Example of a valence line plotted against time, showing areas of high and low valence.

- **Arousal**

Arousal is a measure of a continuous response ranging from one extreme of sleep through intermediate states of drowsiness and the alertness through to frenzied excitement at the other end. This is shown in Figure 4-2, again this example is of the form that could be extracted from a video. Similar to valence, the neutral or "zero" line for the affect-based retrieval system is derived from the average of the arousal values.

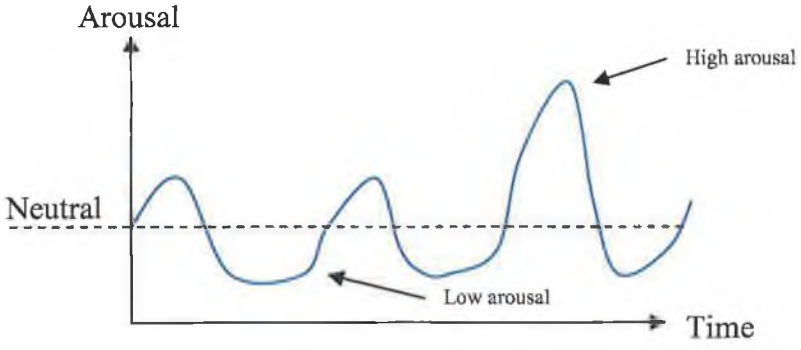


Figure 4-2. Example of an arousal line plotted against time, showing areas of high and low arousal.

- **Dominance**

Dominance (or control) is a measure of a person's perceived control in a situation. It can range from feelings of a total lack of control or submissiveness to the other extreme of complete control or influence over their situation or environment. However, in affective computing, this dimension is generally ignored as there are no known methods of measuring affective features that can be mapped into this dimension. The less discriminative powers of this dimension to separate emotions compared to the other two dimensions is also another contributing factor to its omission [Dietz and Lang, 1999].

Following the cognitive perspective, the human brain is always continually processing and assessing its sensory input and surroundings. The three appraisals or dimensions of valence, arousal and dominance are continually being processed. Therefore the person is said to always be in an emotional state depending on the output of the three dimensions. To visualize this, the three dimensions can be plotted on a three-dimensional coordinate system graph with linear axes ranging from -1 to +1 called the VAD emotion space. Emotional states such as "happy" or "sad" can be said to exist as areas within a certain boundary of coordinates. Therefore when the three dimensions coordinates fall within a boundary region, the person is inferred to be in that particular emotional state.

This gives an intuitive representation of emotions where we can easily infer that the difference between the emotion "happy" and "sad" is that happy falls in the positive regions of arousal and valence, and sad falls in the negative regions. Furthermore, such representation also allows us to understand emotions on a finer level of granularity such as comparing two emotional states "anger" and "rage". While both fall in roughly the same region of positive or high arousal and negative valence, rage can be said to occupy a higher region of arousal than mere anger. This is because rage is often associated with higher levels of activity and violence.

In affective computing, due to the omission of the dominance dimension, this can be plotted as a two-dimensional graph with the same linear axes ranging from -1 to +1 called the VA emotion space. As shown in Figure 4-3, the emotional state "happy" falls in the positive and slightly excited region of the VA space while excitement occupies a

higher arousal region. The difference between "fear" and "angry" is slightly higher arousal but more negative feelings. This way of visualizing emotions as occupying space on a two-dimensional graph means that we can identify the human emotional state by just tracking how engaged and how positive they are. The results of Russell and Mehrabian's [1977] study provided relative values which can be used to populate the VA space with 151 emotion terms that are sufficient to describe a wide range of emotions. These terms are listed in Appendix A (Russell List and Surrey List of Emotional Verbal Labels).

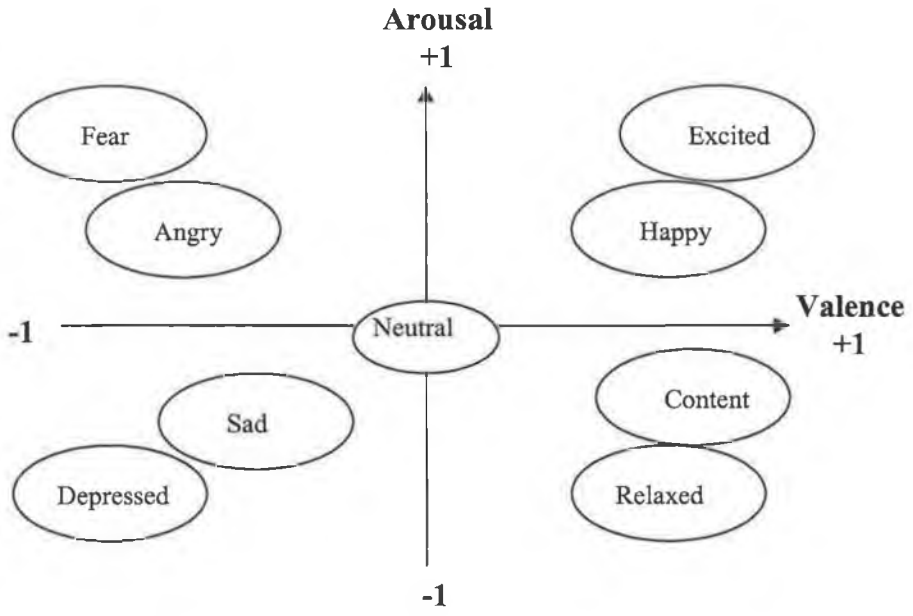


Figure 4-3. A simplified VA emotion space with a few examples of emotions and their distribution.

4.3 Affect Curve

When the detected arousal and valence values are combined over a period of time, this is called the "affect curve" where the line traces the path of the emotions around the 2D VA emotion space over a period of time. Figure 4-4 shows how valence and arousal are combined to form the affect curve.

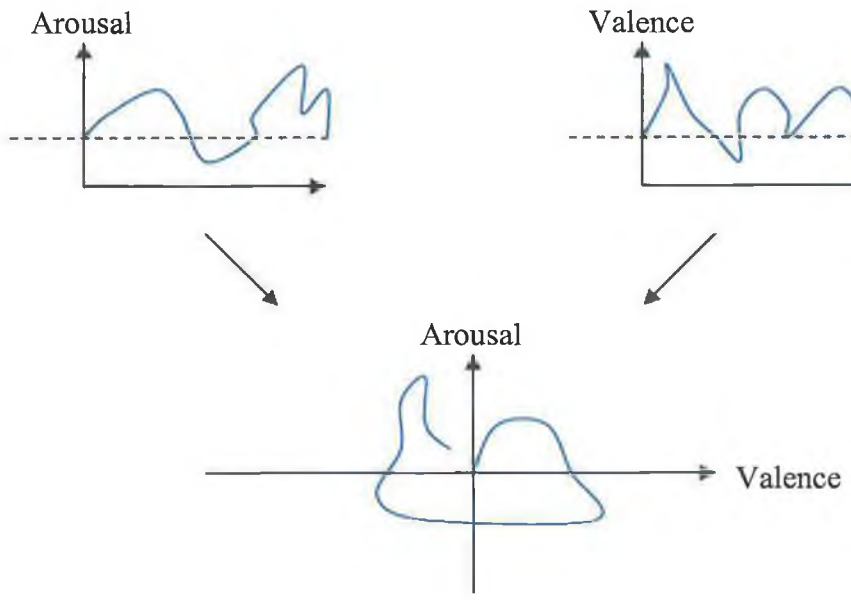


Figure 4-4. When arousal and valence curves are combined, they form the "affect curve".

4.4 Addressing the Criticisms of Affective Computing

In her paper about the challenges facing affective computing, Picard [2003] discusses some criticisms that are leveled at the general area of affective computing. One of the main objections to it is that the range of means and modalities of emotion expression is so broad, with many of these modalities being inaccessible, and many others being too non-differential, it is argued that it is unlikely that collecting the necessary data will be possible or feasible in the near future. The second main criticism is that people's expression of emotion is so idiosyncratic and variable, that there is little hope of accurately recognizing an individual's emotional state from the available data.

Picard addresses the first criticism by referring to work she had conducted that allows at least 8 emotions to be detected using pattern recognition on physiological signals that performed at a level higher than chance [Picard et al., 2001]. Subjects were shown pictures eliciting happiness, sadness, anger, fear, disgust, surprise, platonic love and romantic love. Sensors were used to measure data such as electromyogram, skin conductance or galvanic skin response (GSR), respiration and blood volume pulse over a

long period of time. The results concluded that recognizable physiological signals can be differentiated for the 8 categories of emotions that were investigated.

This is also backed up by other researchers detecting emotions using physiological signals. Lanzetta and Orr [1986] conducted a study differentiating happiness and fear using GSR. They conducted their study by showing subjects slides of facial expressions or administering small electric shocks. They reached a conclusion that fear produced a higher level of arousal and larger skin conductance. Collet et al. [1997] showed neutral and emotionally loaded pictures to subjects in order to elicit emotions from them. The emotions were happiness, surprise, anger, fear, sadness, and disgust. Physiological signals such as skin conductance, skin resistance, skin blood flow, and temperature were recorded and analyzed. Their work successfully distinguished between the emotions by using different combinations of the physiological signals. Nasoz et al. [2003] used emotional film clips to elicit responses from subjects. Various clips with significant emotion were shown and physiological readings were taken using the SenseWear [SenseWear] armband such as GSR, heart rate and skin temperature. They were able to recognize 6 emotions using machine learning algorithms such as sadness, anger, surprise, fear, frustration and amusement.

4.4.1 The Weather Metaphor

These emotion recognition tasks initially might seem similar to other human pattern recognition tasks such as speech recognition. However, words are discrete and therefore the results of a speech recognizer can typically be labeled as either "correct" or "incorrect". Emotions on the other hand, are more ambiguous and therefore the results are harder to define as either "correct" or "incorrect". The results might be "quite correct" or "not as correct", which leads to the second criticism that people's expression of emotion is so idiosyncratic and variable, that there is little hope of accurately recognizing an individual's emotional state from the available data. To address this, emotions can be described using the weather as a metaphor:

> The term emotion refers to relations among external incentives, thoughts and changes in internal feelings, as weather is a superordinate term for the changing relations among wind velocity, humidity, temperature, barometric pressure and form or precipitation. Occasionally, a unique combination of these meteorological qualities creates a storm, a tornado, a blizzard, or a hurricane—events that are analogous to the temporary but intense emotions of fear, joy, excitement, disgust, or anger. But wind, temperature, and humidity vary continually without producing such extreme combinations. Thus meteorologists do not ask what weather means, but determine the relations among the measurable qualities and later name whatever coherences they discover...[Kagan, 1984]

Thus Picard [2003] reasoned that it would be feasible to build a system for detecting emotion by measuring the equivalents (such as human physiological signals) of temperature, pressure, and humidity as found in a weather prediction system. The different streams of data might have patterns that could be combined, thus detecting the emotional equivalents of a tornado or blizzard.

In the case of the weather, tornadoes and blizzards are extreme states that exist when certain unique combinations or criteria are met. These can be compared to extreme human emotional states such as intense fear that could be detected by extreme fluctuations in physiological signals such as sweating or rapid heart beats.

However, most states of weather do not even have names. For in-between states of weather, adjectives such as "partly cloudy" or "nice day" are often used to describe the quality of the weather. This mirrors the usage of adjectives such as "happy" or "slightly nervous" to describe emotions. Depending on its intended purpose, such descriptions are perfectly acceptable for use, just as in a weather report.

Thus, directly detecting a person's emotion might be extremely difficult without new knowledge being made in psychology and the human cognitive process. However, detecting the features which can influence human emotions (affective features) can infer the underlying emotional state.

4.5 Examples of Affective Computing Applications

Affective computing allows us to include emotions into computing processes in order to provide unique and useful user-centric applications that cannot be realized otherwise.

One unique example application is described by Inanoglu and Caneel [2005]. They built a system called "Emotive Alert" for sorting audio voicemail according to various affective features such as urgency, formality, happy or sad. They demonstrated an application of their system on a mobile phone which is able receive voicemails, and then classify and sort them according to eight emotions models. Voicemails by callers that sound more urgent can then be sorted and listed first for immediate attention. They trained a Hidden Markov Model (HMM) based classifier on extracted audio features such as audio pitch, loudness and speaking rate to classify voicemails between the eight emotions.

Another potential application of affective computing is in e-mail and instant text messaging clients that could automatically detect their user's emotional state while they are typing an e-mail and generate emoticons on the fly. A system could also be built around summarizing huge volumes of text according to its affective content for use in a text search engine. Liu et al. [2003] presented an approach that categorizes text documents into one of the six emotions categories (happy, sad, angry, fearful, disgusted, surprised and neutral) and presented the results by associating the text and their emotional content with colours. Classification is done by using four linguistic models of text built from Open Mind Commonsense, a large corpus of commonsense knowledge about the affective quality of things, actions, and events. The corpus was built using commonsense knowledge supplied manually on a large scale by human contributors. The affective content in the text documents are then represented with colours, because they are visually distinct and have emotive information that user can quickly identify.

Virtual avatars in videoconferencing and gaming sessions can be augmented with emotional expressions to allow people to express themselves. Current research into face detection based on Ekman and Friesen's [1978] work on Facial Action Coding System (FACS) can be modified to automatically look for these affective cues and detect their associated emotions for display. Their pioneering work produced a system for describing

"all visually distinguishable facial movements", called the Facial Action Coding System or FACS. This relies on movements of groups of facial muscles that are measured and defined as "Action Units" that can be used to analyze facial expressions. Essa and Pentland [1997] demonstrated their facial expression recognizer that was able to recognize 4 types of emotions (Happiness, Surprise, Anger, Disgust). Their approach was based on tracking and recognizing facial motions from videos of faces. The results of their recognizer, which can be as high as 98% accurate purely from video data, demonstrate how virtual avatars in videoconferencing and gaming can be given emotional expressions without the need for invasive physiological sensors.

A unique investigation into the potential that affective computing offers is the work done by Shugrina et al. [2006]. Their "Empathic Painting" system utilizes face detection techniques based on FACS to paint a picture based on a viewer's emotion. A graphic painting is rendered, distorted, and modified based on a video feed of the viewer's face to reflect their emotional expression. Such work offers new potential for interactive user interaction such as an installation in an art gallery.

4.6 Affective Computing for Content-Based Video Retrieval

Affective features offer a new dimension of information that can be used for content-based classification and retrieval. The ability to detect emotional content is highly useful especially for video and audio files as their temporal nature can convey a lot of emotional content. This emotional content is naturally interpreted by human viewers and often plays a vital part in understanding the video content itself. Thus it would be useful from a CBVR perspective to offer users a method to discover and interact with emotional content to allow efficient access to their information need.

As discussed in Section 3.3 (Bridging the Semantic Gap with Affective Computing), Hauptmann [2005], in his review of the lessons of CBVR systems, noted that one way to bridge the semantic gap between user needs and detected features is to take features and combine them to infer some sort of high-level ontology that non-expert users can relate to. He favoured this method because this splits the semantic gap problem into two:

- Mapping low-level features to immediate semantic concepts
- Mapping these semantic concepts to user needs

He identified affective computing as a promising avenue of research to tackle this problem due to its unique user-centric focus. Recent work in affective computing is focused on the first problem. While emotions cannot be clearly defined, affective features that contribute and influences emotions can be. This is why the VA emotion space is such an important foundation for affective computing. The VA emotion space breaks apart the problem of describing emotion into the two dimensions of valence and arousal. Valence and arousal can be measured based on features in multimedia data. This provides a framework for modeling valence and arousal on low-level features. This thesis extends this work beyond addressing the first problem, to exploration of a solution to the second problem as well.

4.7 Using Films as Video Data in Emotion-Related Research

Related work in affective computing can generally be grouped into face detection, human voice affect detection and music analysis, but applying it into the domain of commercially created video, such as films is a new direction. Hollywood-style films constitute a large portion of the entertainment industry where almost 4500 films are released yearly around the world, corresponding to approximately 9000 hours of video data. This represents a huge potential resource for studying the relationship between humans and emotions.

Films contain the most expressive of human acting and film-making techniques that are designed to elicit an emotional response from viewers through the art of storytelling. This medium is one of the more mature and popular forms of human entertainment and leverages decades of film-making knowledge that is designed to capture the heart and imagination of its viewers. This makes it one of the most affect-laden media in video and it is a great source of data for affective features. An increasing

amount of research in affective computing is just beginning to mine information from this domain [Rasheed et al., 2005].

There exists a large body of knowledge and rules accumulated in film theory, film language or cinematography that guides film-makers in making not only movies but also documentaries, sitcoms and other entertainment shows. This makes the results of analysis performed on movies applicable to other domains of entertainment videos as well. Also called "the grammar of film", these are not necessarily set rules which film-makers take, as it is up to their creativity to portray and tell a story. Some intentionally break these "rules" to come up with new perspectives to story-telling. Even so, these are common techniques designed to elicit an emotional response from viewers that are frequently used and can be detected by computers as affective features. Much of the related work that is described in the next section (Section 4.8 Related Work in Detecting Emotions from Films) relies on some simple assumptions related to these rules.

- **Editing**

Scene editing refers to how editors or directors cut up and mix recorded scenes to tell a story. For example, horror movie directors sometimes use long periods of silent and slow scenes to build up tension before following it with a loud bang to surprise the audience. A jump cut (or a shot cut), which is an abrupt switch from one scene to another, can be used to make a dramatic point, or indicate a change of scene. An increase in the rate of these cuts usually indicates an increase in activity.

- **Camera Techniques**

Camera movement refers to how the camera is being manipulated to frame the event that is happening in the film. There are a lot of types of camera movements, such as panning left or right, zooming in and out, tilting up and down or tracking the movement of an object or person. An increase in camera motion can also indicate an increase in activity.

- **Colour and Lighting**

Colour and lighting can be used to manipulate the viewer's emotions and attention. For example, a flash of bright light can often be used to indicate a near-death scene or religious experience. In the series of movies such as The Matrix, a green tinge or filter is often applied to specific scenes to suggest artificiality (the green colour as found on early monochrome monitors) and link them together so that viewers are able to perceive them as being a different but coherent artificial world when they see the green tinge.

- **Sound**

Sound plays an important role in determining the emotional impact of a scene as well and is used to great effect by movie-makers to elicit an emotional response. For example, horror movie directors typically use long periods of silence to build up tension before scaring the audience with a loud and high-pitched sound. There are also more subtle uses of different types of music to convey a different mood or feeling such as using orchestral music to signify grandness or a violin solo to indicate sadness.

4.8 Related Work in Detecting Emotions from Films

Hanjalic and Xu [2003] carried out pioneering work using the VA emotion space to model low-level video features onto the valence and arousal dimension. Their work showed that valence and arousal can be effectively modeled by low-level video features such as motion, shot cut rates, audio energy and audio pitch by utilizing basic knowledge of cinematographic techniques.

According to Hanjalic and Xu, motion is one of the more extensively investigated visual features in the context of video content analysis. This is also observed by film theorists who contend that motion is highly expressive and is able to evoke strong emotional responses in viewers. Work by Detenber et al. [1997] and Simons et al. [1999] investigated the influence of camera and object motion on emotional responses of humans and concluded that an increase of motion intensity on the screen causes an increase in arousal. In addition, the type of emotion (valence) was found to be

independent of motion; if the mood of a person was "positive" or "negative" while watching a still picture, the mood will not change if motion is introduced within that picture.

In addition, varying shot length or shot cut rate can also be a good example of an editing effect that can be put in relation to the affective content of the video. Editing plays such as in important role that directors often manipulate this to achieve certain effects. Typically shorter shot lengths are perceived as being more action-oriented, a high tempo of action, or stressed, accented movements. The opposite of this, which is longer shot lengths, are typically used to de-accentuate an action. Outside of highly scripted videos, live broadcasts of sports video still holds this "rule" to be true. For example, in a soccer match in which most of the match is broadcast from one camera that covers the whole field. When there are interesting events such as the scoring of a goal, the director in the control room will try to cover much of the action by switching between cameras, resulting in higher shot cuts. Therefore an increase in shot cuts can be related to higher arousal and vice versa.

Based on the results of Murray and Arnott [1993], Slaney and McRoberts [2003] and conclusions by Picard [1997], various vocal effects present in the sound track of a video may bear broad relations to the affective content of the video. In terms of affect dimensions, loudness is often related to arousal while pitch-related features are commonly related to valence. However, as acknowledged by Picard [1997], detecting valence can be more difficult than arousal as valence information is more subtly conveyed.

Hanjalic and Xu combined the valence and arousal dimensions to plot the affect curve which shows broadly the trends of emotional development in a piece of video. Kok [2006] proposed the use colour saturation and scene brightness as extra visual features towards modeling the valence component of the affect curves. His work was motivated by the work of [Valdez and Mehrabian, 1994], which discovered that the saturation and brightness had strong and consistent effects, therefore the more saturated and brighter the colour, the more positive the perceived emotion. For the affect-based retrieval system presented in this thesis, the results of Kok's work were implemented to extend Hanjalic and Xu's model of valence.

Rasheed et al.'s [2005] work of classifying films into four categories (comedies, action, drama, and horror films) examines the use of low-level video features that take into account some knowledge of ubiquitous cinematic practices. In the course of their work, they identified shot cuts, colour, motion, and lighting as useful and relevant visual features, grounded in cinematic practices, that can be used for building content-based video retrieval systems for classifying films into different genres.

Hang's [2003] work on affective content detection relies on the VA emotion space to model low-level features to 4 emotion categories as shown in Figure 4-5. Using video data, he used a HMM-based classifier to discriminate between 4 emotion categories using low-level features such as colour, motion and shot cut rates.

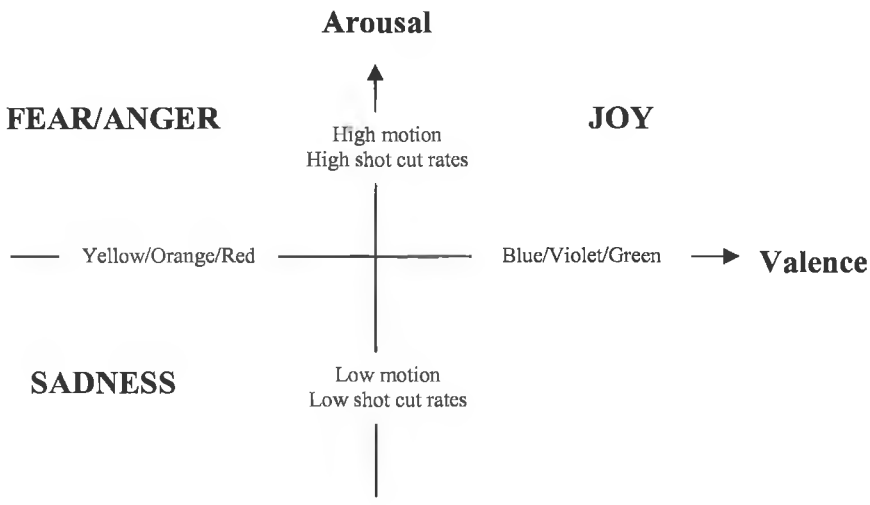


Figure 4-5. Using the VA emotion space to model emotions by [Hang, 2003]. Low-level features such as colour, motion and shot cut rates can be associated with the 4 emotions Fear, Anger, Joy and Sadness.

Salway and Graham [2003] presented a method to extract information about a character's emotion in a film based on textual descriptions of the content. Their approach used "audio descriptions" for the visually impaired that are contained in the VHS and DVD film releases. These audio descriptions were identified as being a valuable source

of information for improving access to higher levels of multimedia semantics relating to the narrative structure of films. The following is an example of one of these descriptions along with their time-codes describing scenes in a film (The English Patient) and where it occurred:

[11:43] Hanna passes Jan some banknotes
[11:55] Laughing, Jan falls back into her seat
[12:01] An explosion on the road ahead
[12:08] The jeep has hit a mine

These descriptions were analyzed for possible descriptions of emotions in order to classify them into 22 types of emotion. Using the online dictionary WordNet [WordNet], each type of emotion such as "Joy" is expanded into keywords such as "euphoria", "jolly", "happy", and "pleased", therefore allowing audio descriptions where these keywords occur to be classified into one of the 22 emotion types. Depending on where the keywords occur, a plot of emotion words can be generated as shown in Figure 4-6.

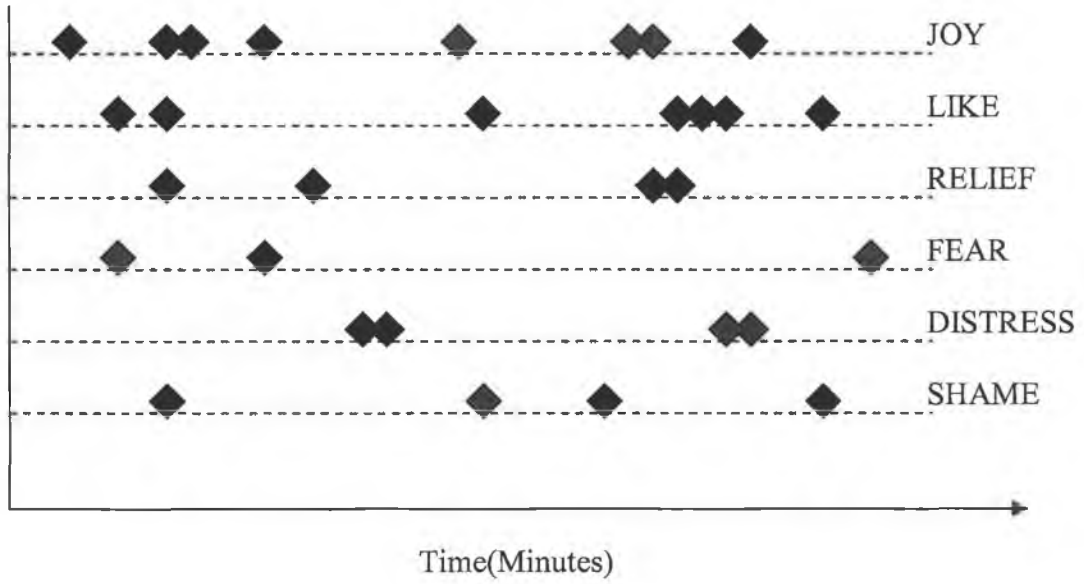


Figure 4-6. Example of a plot of emotion tokens found in the audio descriptions of a film. Taken from [Salway and Graham, 2003].

While encouraging progress is being made with mapping low-level visual and audio features to clear semantic structures, in later work, Salway et al. [2005] in an effort to bridge that semantic gap, analyzed screenplays alongside audio descriptions that describe films in order to identify frequently occurring phrases such as "looks at" and "smiles at". These phrases occur frequently because they describe actions that are important story-telling elements in the film. When these occurring phrases are plotted against time on a graph as shown in Figure 4-7, similar to Figure 4-6, they observe a clustering effect for some of the phrases. These clusters can be observed in Figure 4-7 for phrase "door" and "looks at". This clustering could indicate dramatically important sequences or developments in the story telling. By using such a visualization tool, they can identify important scenes and attempt to build a system and ontology for navigating film content.

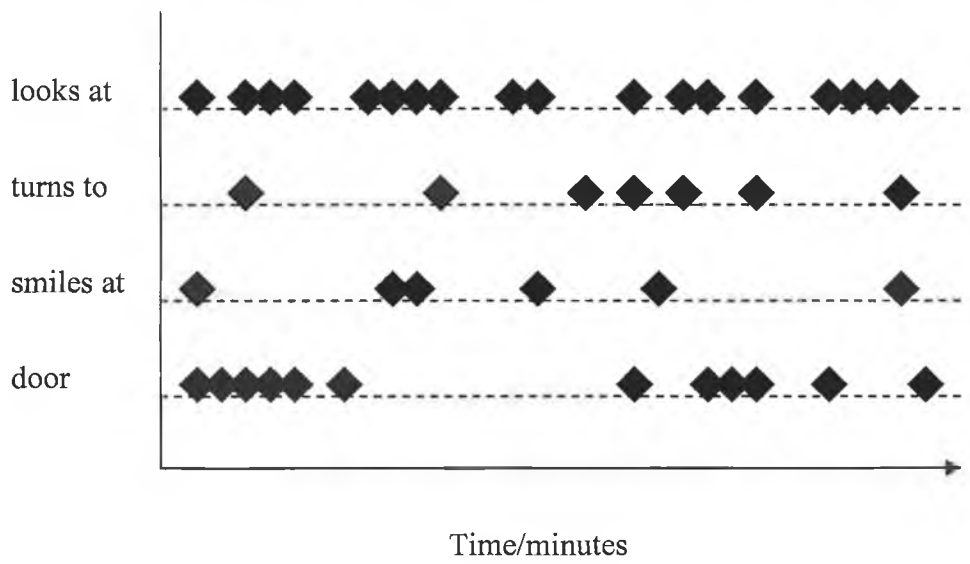


Figure 4-7. Example of a plot of phrases occurring in the description of a film. Taken from [Salway et al., 2005].

4.9 Summary

This chapter introduces affective computing and how it can be applied to CBVR. Psychological studies into the human cognitive process conclude that emotions play an important role in human perception, decision-making, and interaction. People display emotions even when interacting with machines that have no emotions. Intelligent systems that can detect emotions therefore can better serve their users by analyzing and reacting appropriately based on the user's emotional state.

In addition, video has a rich amount of emotional content that can be used for video annotation and retrieval. Emotion can be hard to define, therefore recent work in affective computing uses the VAD emotion space to describe emotions with easier-to-define components. We can then infer the emotional or affective state of a person by detecting these components separately and combining them into an "affect curve". For video retrieval applications, knowledge of "film grammar" typically used by film makers can be used to detect these components as well. This enables the affective states of a video to be detected and used for video retrieval purposes. The chapter ends with some examples of work in affective computing for video and film analysis.

Chapter 5 Information Retrieval and Document Matching

In order to search for multimedia documents annotated with textual labels of relevant affective features, well established text retrieval methods are used. This chapter presents an overview of information retrieval (IR) and the Okapi BM25 model used in the affect-based video indexing and retrieval system as presented in this thesis. This chapter also introduces the WordNet ontology and word similarity metrics used to implement open-vocabulary affect-based retrieval in this thesis.

5.1 Introduction

Conventional research has mainly focused on computer-readable text. The most visible results of such research are commercial web search engines which play a vital role in the Internet such as Google and Yahoo. These web search engines are usually the first stop for people looking for information on the Internet. The sheer amount of information on the Internet makes it almost impossible to locate specific documents without the use of such search engines.

The classic IR problem is to locate the desired or relevant text documents using a search query consisting of a number of keywords. Matching the keywords from the query with the keywords within the document retrieves the relevant documents. If a document has a higher amount of keywords that match the query then it is regarded as being more important or relevant than other documents that have less. Using this principle, documents can be ranked according to relevance for users to browse. Powerful algorithms have been developed which combine a number of factors to improve retrieval effectiveness. The methods have been thoroughly evaluated for the TREC evaluations [Harman, 1997]. The following sections describe the basis of IR and the elements of one of the most used, tested and effective methods for text retrieval, the Okapi BM25 IR model [Robertson and Sparck Jones, 1997].

5.2 Terms and Matching

The retrieval features are referred to as search terms. They may be full words or features reduced to stem or root terms to encourage matching using a stemming algorithm. Terms are generally stems or root words rather than full words in order to avoid missing matches through word variations. Stemming can be achieved by truncating user request words, in order to match any index words. In this thesis, for simplicity words are used as the search terms.

The request is taken as an unstructured list of terms. If the terms are not weighted, the results can be ranked by the number of matching terms between the request and each document. However, performance can be improved by using weights, as described in the next section. Here term weights are used, a matching score is formed for each document by summing the weight of terms appearing in the search query and the document.

5.3 Term Weighting

By assigning weights to terms, performance can be improved considerably. The idea behind term weighting is selectivity. A good term is the term that can pick the relevant documents from the non-relevant documents. Term weighting is designed to assign higher weights to terms which exhibit characteristics of providing high selectivity for relevant documents. There are three different sources of weighting data in classical IR term weighting:

- **Collection Frequency**

Terms that occur in only a few documents are often more valuable than ones that occur in many. Collection frequency weights are defined as follows, for a term t(i),

n = the number of documents term t(i) occurs in
N = the number of documents in the collection

The collection frequency weight (CFW) for a term is given by,

$$CFW\ (i) = \log N - \log n$$

- **Term Frequency**

The second source of weighting is a term's within-document frequency, where the more often a term occurs in a document, the more likely it is to be important for that document. Therefore while a term's collection frequency is the same for any document, its document frequency varies. The term frequency (TF) for term t(i) in document d(j) is,

$$TF\ (i,j) = \text{the number of occurrences of term } t(i) \text{ in document } d(j)$$

However, term frequency should not be used just as it stands as a weighting factor, but must in general be related to the remaining source of information about documents.

- **Document Length**

The third input to weighting is the length of a document. A term that occurs the same number of times in a short document and in a long one is likely to be more valuable for the former since on average tf(i,j) will increase with document length. This means that longer documents will in general have higher weights. Since the document matching score is formed by summing the matching term between the query and the document, longer documents will tend to score higher, i.e. there will be a bias in favor of longer documents in the retrieved ranked list. To overcome this bias a document length compensation component is often incorporated in the term weights.

The use of the document length described below normalizes the document length by the length of an average document, using the normalized document length NDL(j).

$$DL(j) = \text{the total of term occurrences in document } d(j)$$

$$NDL(j) = \frac{DL(j)}{\text{Average DL for all documents in the collection}}$$

This has the advantage that the units in which DL is counted do not matter much. A very simple measure such as number of characters in d(j) can be quite adequate as a substitute for number of term occurrences. In this thesis the number of annotated affect labels for each document is used as the document length.

5.4 Combining the Evidence

The three kinds of weighting data for each term can to be combined together to form a combined term weight for each term on each document. There are various formulae for this combination, one of the most popular and consistently effective is the Okapi BM25 IR model [Robertson and Sparck Jones, 1997]. The BM25 term weight is defined as follows, for a term t(i), in a document d(j), the BM25 combined weight CW(i,j) is,

$$CW(i,j) = \frac{CFW(i) \times TF(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times (NDL(j)))) + TF(i,j)}$$

K1 and b are tuning constants. The formula ensures that the effect of term frequency is not too strong and for a term occurring once in a document of average length, the weight is just CFW(i) as described earlier. The overall matching score for a document d(j) is simply the sum of the weights of the query terms present in the document. Documents are ranked in descending order of their matching scores for presentation to the user.

The tuning constant K1 modifies the extent of the influence of term frequency. Ideally, it should be set after systematic trials on the particular collection of documents. The constant b, which ranges between 0 and 1, modifies the effect of document length. If b=1 the assumption is that documents are long simply because they are repetitive, while if b=0 the assumption is that they are long because they are multi topic. Thus setting b towards 1, such as b=.75, will reduce the effect of term frequency on the grounds that it is primarily attributable to verbosity. If b=0, there is no length adjustment effect, so greater length counts for more, on the assumption that it is not predominantly attributable to verbosity.

The BM25 IR model is used in the affect-based retrieval system as presented in this thesis to perform video retrieval. Users input a typed text query and the BM25 IR model is then used to match the query with the system-detected emotional verbal labels and retrieve the relevant videos.

5.5 WordNet and Relatedness Measures

In standard text IR, documents and queries both use an open vocabulary and the success of an IR system relies on there being a good match between words appearing in relevant documents and the submitted search request. The verbal labeling of the affective features described in Section 4.2 (VAD Emotion Space) is limited to 151 words. The 151 words or labels are listed in Appendix A (Russell List and Surrey List of Emotional Verbal Labels). These words may not be those chosen by a user in a request, preferring other ones instead. Mismatch between the terms used by the user and the 151 affect labels will mean that the documents will not be retrieved. In order to address this problem a method is needed to gain a measure of similarity between the query words entered by the user and the document index words. This thesis proposes a novel solution to use freely available software WordNet and WordNet::Similarity to enable user queries to be mapped onto the limited set of emotional labels detected by the affect-based indexing and retrieval system (Section 6.9 Open-vocabulary Query Mapping). This section will describe the electronic lexical dictionary WordNet and WordNet::Similarity, a freely available software package for measuring relatedness distance between two words using

the WordNet dictionary. Because the two software packages are used in the development of the affect-based video indexing and retrieval system, some understanding of some key concepts and how the software works is presented here.

5.6 WordNet

WordNet is a freely-available electronic lexical database/dictionary, developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller [WordNet]. Inspired by psycholinguistic theories of human lexical memory, the dictionary consists of nouns, verbs, adjectives and adverbs that have been organized into related concepts called synonym sets or synsets. For example, "car, auto, automobile, machine, motorcar" is a synset that represents a concept. Each synset is defined by a gloss (description), which in this example, would be: "4-wheeled motor vehicle; usually propelled by an internal combustion engine". These descriptions provide the definition or example sentences which identify separate synsets.

The nouns, verbs, adjectives and adverbs in WordNet have different relations to represent and connect similar concepts together. These relations create a network where related concepts can be identified by their relative distance from each other. These relations vary based on the type of word and include; synonymy, antonymy, hyponymy (is-a), and meronymy (part-of). Due to the way the words are organized into synsets, WordNet can potentially be a valuable tool for measuring semantic similarities between words that can be applied to query words and detected keywords in a video retrieval system.

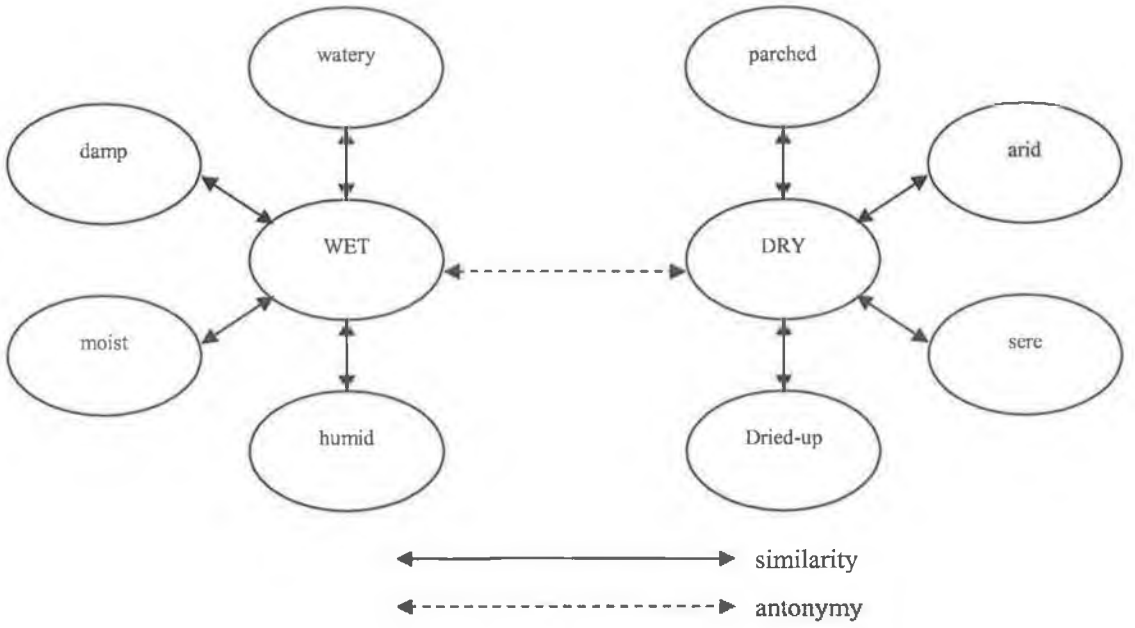


Figure 5-1. An example of the relations between two adjectives "wet" and "dry" and other related words in WordNet.

As an example shown in Figure 5-1, the two adjectives "wet" and "dry" are connected by an antonym relationship. Other words in the WordNet ontology such as "damp", "watery", "moist", and "humid" are connected to "wet" with a similarity or synonymy relationship. The relations therefore can be used determine how each word is related to each other. For each word in WordNet, there are four kinds of relations that connect each word to others in WordNet's lexical structure. The 4 relations are:

- **Synonymy**

Synonymy refers to the similarity of meaning. This lexical relation is the most important relation for WordNet because the ability to judge the relation between word forms is a prerequisite for the representation of meanings in a lexical matrix. One possible definition is that two expressions are synonymous if the substitution of one for the other never changes the truth value of the sentence. But by this definition, such cases are rare if they exist at all, so a weakened version of this definition would be to make synonymy relative to a context. One can infer that two words have a synonymous relation if they are

interchangeable in a sentence without modifying the truth in its context. For example, the substitution of the word {plank} for {board} in a carpentry context will seldom alter the truth. Such a definition makes it necessary for WordNet to be partitioned into nouns, verbs, adjectives and adverbs. This is because nouns express nominal concepts, verbs express verbal concepts, and modifiers provide a way to qualify the concepts, making substitution between them impossible without altering their context.

- **Antonymy**

Although antonyms might seem to be a simple symmetric relation, they are actually quite complex. Antonym in itself means a word of opposite meaning, but in WordNet the antonym of a word x is sometimes not-x but not always. For example, {rich} and {poor} are antonyms, but to say that someone is not rich doesn't mean that they must be poor. So antonym is a lexical relation between word forms, not a semantic relation between word meanings. Antonyms provide a central organizing principle for the adjectives and adverbs in WordNet.

- **Hyponymy/Hypernymy (is-a relation)**

Hyponymy, also known as an is-a relation, is a semantic relation unlike synonymy and antonymy. A concept represented by the synset {x,x',...} is said to be a hyponym of the concept represented by the synset {y,y',...} if native English speakers accept sentences constructed from such frames as 'an x is a (kind of) y'. So for example, a {maple} is a (kind of) of {tree}, and {tree} is a (kind of) of {plant}. Because of such a subset/superset relation, these relations generate a hierarchical semantic structure with a single super ordinate. As one traverses down the structure of nodes, the hyponym's concept becomes more specific or topical, inheriting all the features of the generic concept. This convention provides the central organizing principle for the nouns and verbs in WordNet.

- **Meronymy/Holonymy (part-of relation)**

Meronymy, also known as a part-of relation, is a semantic relation as well, similar to hyponymy. A concept represented by the synset {x,x',...} is a meronym of a concept represented by the synset {y,y',...} if native English speakers accept sentences constructed from such frames as 'a y has an x (as a part of) or an x is a part of y'. So for example, {arm} is a part of {body}.

5.7 WordNet::Similarity

WordNet::Similarity is a freely-available software package implemented as modules in the Perl programming language [WordNet::Similarity]. The semantic relatedness measures that are implemented in this software by Pedersen et al. [2004] include those described by Leacock & Chodorow [1998], Resnik [1995], and Hirst & St-Onge [1998]. This software makes it possible to measure the semantic similarity or relatedness between a pair of words using the WordNet database.

5.7.1 Measures of Similarity

Measures of similarity quantify how much two concepts are alike based on the information contained in WordNet's relations or hierarchy. Because of the unique feature of WordNet in organizing words into related concepts, two words can be said to be similar if counting the distance of the relations or hierarchy results in a small distance. Extending this further, a similarity measure can be derived by counting the distance from one word to another. For example, {automobile} can be considered more similar to a {boat} than a {tree}, if {automobile} and {boat} share {vehicle} as a common ancestor in their is-a hierarchy, resulting in a shorter distance between the two.

The similarity measures in WordNet::Similarity are based on counting path lengths between concepts, utilizing the is-a relation. WordNet is well suited for similarity measures because it organizes nouns and verbs into hierarchies of is-a relations. Two path counting algorithms are:

- **Leacock and Chodorow**

Presented in [Leacock and Chodorow, 1998], this algorithm finds the shortest path between two concepts and scales this value by the maximum path length D in the is-a hierarchy in which they occur. Thus,

$$sim_{LC}(c1, c2) = - \log \frac{length(c1, c2)}{2D}$$

where $sim_{LC}(c1,c2)$ is the similarity measure between two words c1 and c2, with length(c1, c2) being the shortest path between two synsets for those words.

- **Resnik**

Resnik's approach is based on information content, which uses knowledge of a corpus to derive a similarity measure [Resnik, 1995]. Guided by the intuition that similarity between a pair of words may be judged by "the extent to which they share information", Resnik defined the similarity between two words in WordNet to be the information content of their lowest super-ordinate (most specific common subsumer) lso(c1, c2):

$$sim_R(c1, c2) = - \log p(lso(c1,c2))$$

where $sim_R(c1,c2)$ is the similarity measure and p(c) is the probability of encountering an instance of a synset c in some specific corpus.

5.7.2 Measures of Relatedness

WordNet is well suited for making similarity measures due to their is-a relations hierarchy. However, is-a relations in WordNet do not cross part of speech boundaries, so such measures are limited to judging between noun pairs and verb pairs. The affect-based system presented in this thesis relies on adjective words that describe emotions such as "happy", "joyful" and "sad" in order to describe affective states.

Adjectives and adverbs in WordNet are not organized into is-a hierarchies. Although they don't have is-a hierarchies, they can still be related in other ways beyond being similar to each other. They are organized by different relations such as antonyms and part-of relations, therefore called measures of relatedness instead. For example, a {wheel} is a part of a {car}, and {night} is the opposite of {day}. {Happy} have antonyms such as {unhappy}, and synonyms such as {blissful} and {bright}. In addition, more information about them can be retrieved from their accompanying gloss as well. Measures of relatedness are based on these additional sources of information, and as such can be applied to a wider range of concept pairs.

Almost all similarity measure algorithms in WordNet::Similarity rely on using is-a relations as a measure of similarity, which do not exist for adjectives, therefore the measure of relatedness proposed by Hirst and St-Onge (1998) is used for measuring word relatedness in the affect-based system.

- **Hirst and St-Onge**

Hirst and St-Onge's [1998] idea of semantic relatedness is that two words are semantically close if their WordNet synsets are connected by a path that is not too long and "does not change direction too often". In this case, each of the relations is classified as horizontal, upward or downward. Upward relations connect more specific concepts to more general ones, while downward relations join more general concepts to more specific ones. The is-a relation is an example of an upward relation while the part-of relation is a downward relation. Antonyms are horizontal relations which maintain the same level of specificity. Therefore, this measure of relatedness relies on finding a path between two

words through any relations including antonyms which are needed for the adjectives used in the affect-based video indexing and retrieval system presented here. The strength of the relationship between word c1 and c2 is,

$$rel_{HS}(c1, c2) = C - \text{path length} - k \times d$$

where $rel_{HS}(c1,c2)$ is the relatedness measure of words c1 and c2, d is the number of changes of direction in the path, and C and k are tuning constants set to C=8 and k=1 as described by Patwardhan et al. [2003].

The affect-based retrieval system that is presented in this thesis uses this measure of relatedness to compare how similar two words are. Therefore relatedness and similarity is used interchangeably for the rest of this thesis. During the IR stage of the system, each query word input by the user is compared to each of the system's pre-defined emotional verbal labels using the Hirst and St-Onge's measure in WordNet::Similarity to derive the relatedness (or similarity) scores. The system's pre-defined labels with the highest score is then said to be the most related to the user's query word.

5.8 Summary

This chapter has described the information retrieval and document matching methods that are used in the affect-based retrieval system that is presented in this thesis. The Okapi BM25 IR model is proven to be effective in retrieval applications and is used in this thesis. Therefore some concepts of text IR and the Okapi BM25 IR model were presented in the beginning part of this chapter.

This was then followed by an introduction to WordNet and an explanation of the relationships between words that are used. The chapter also introduces WordNet::Similarity and the measures of similarity and relatedness along with some of its algorithms. In order to allow the affect-based retrieval system presented in this thesis

to extend it's vocabulary beyond a fixed set of words with pre-defined affective components, the WordNet ontology and the WordNet::Similarity measures of relatedness is used to map unknown words onto the system's known words. However, the is-a relationship is not present in WordNet for adjectives that form the words that are used by the affect-based retrieval system, therefore the Hirst and St-Onge measure of relatedness is used.

Chapter 6 Design of a Novel Affect-Based Video Indexing and Retrieval System

In order to explore the potential of affective features for CBVR-related applications, a system was built that enables users to retrieve videos based on their affective content. This chapter presents the design of this novel affect-based video indexing features and their integration into a retrieval system along with the details of its implementation.

6.1 Introduction

The aim here is to build an automatic video indexing and retrieval system that enables users to browse through videos according to their affective content. Affective content refers to the stimuli that are being invoked to try to influence a viewer into an emotional state. It is therefore far more consistent to try to index and label the affective features of a piece of video instead of trying to detect the viewer's emotion directly (Section 4.1 Affective Computing: Introduction).

The prototype system was developed using Matlab version 7 release 14, a high level language and interactive environment that comes with an extensive range of library functions called toolboxes designed for various applications. Two toolboxes were used: the Signal Processing Toolbox and Image Processing Toolbox, to eliminate the need to re-code commonly-used functions in the signal processing and image processing community. In addition to the toolboxes, some script files containing block matching algorithms for motion estimation was also downloaded from the Matlab website [Matlab].

In addition to Matlab, two other software packages were used; the WordNet lexical dictionary [WordNet] and WordNet::Similarity, a collection of functions to measure word relatedness [WordNet::Similarity]. WordNet::Similarity is implemented as a Perl module that can be invoked to interface with the WordNet dictionary. Matlab supports calling Perl scripts from within its development environment, therefore almost all of the development of the system was done on Matlab.

The affective retrieval system consists of Matlab scripts that are able to take a text query input from a standard text file and return a ranked list of retrieval results after processing. Still within Matlab, using the list of returned results, users can then play back specific video segments that are deemed by the system to contain the emotions as requested in the query. For video playback, Matlab invokes the VideoLAN (VLC) media player which is a freely available media player that can play various video file formats [VLC].

6.2 System Considerations

Entry of a textual query enables users to search for videos by typing in a text query such as "happy and elated", and be presented with a sorted list of videos that have been detected as fulfilling the user's query. This approach is taken for a number of reasons:

- As mentioned earlier, affective computing is aimed at providing user-centered applications that facilitate their information needs. A simple and familiar interface for users to use is therefore required. All popular web search engines use a textual query interface and present users with the results in the form of a sorted list. This makes it one of the most widely used forms of interaction and will be familiar to almost everyone regardless of computer knowledge; therefore the system presented here uses this approach as well.
- Emotions can be hard to describe. For users of a system that relies on emotional content for indexing and retrieval, it is important that users are able to express their queries in a natural, simple and descriptive way. The use of a text query allows users to input query words that gives them the flexibility to describe emotions using a variety of words.
- In addition, analysis of feedback from the human subjects watching films in the CDVPlex project [Rothwell et al., 2006] found that users typically use a wide range of words to describe the emotional content found in films. While there are a lot of recurring simple words such as "happy", "sad", and "fear", this is usually augmented with more descriptive words such as "dynamic", "angst", and "triumphant". Using the list of 151

words as described in Section 4.2 (VAD Emotion Space) and listed in Appendix A (Russell List and Surrey List of Emotional Verbal Labels), the system can detect an adequate amount of words or emotional labels, however there will be query words that do not exist in the system. In order to bridge this gap between system-detected labels and user queries i.e., not to restrict a user's query, an open-vocabulary query approach is implemented using the WordNet lexical dictionary to accommodate for words that does not exist in the system.

- Textual queries also allow the system to leverage on established research in text information retrieval (IR) as discussed in Chapter 5. Simple and proven IR methods such as BM25, ensures that the query system's performance in large-scale retrieval tasks is not compromised [Robertson and Sparck Jones, 1997].
- The affective features used in this system are extracted purely from the video and audio data. Recent work in affective features is concentrated on physiological signals or recognizing human facial gestures to detect emotion, but for indexing and retrieval applications, it would be intrusive and impractical to obtain such data for indexing video. In addition, human emotion varies depending on the individual; therefore one person might have a different response than another watching the same piece of video. Thus detecting the affective content purely from video and audio data provides consistency and practicality. This is because the affective information contained in the video does not change depending on the viewer, thus giving consistency to the annotating labels. In addition, through human user evaluations, consistency in retrieval results can also be attained if a large number of users rated the system's retrieval performance favorably.

6.3 System Overview

The system is structured as shown in Figure 6-1:

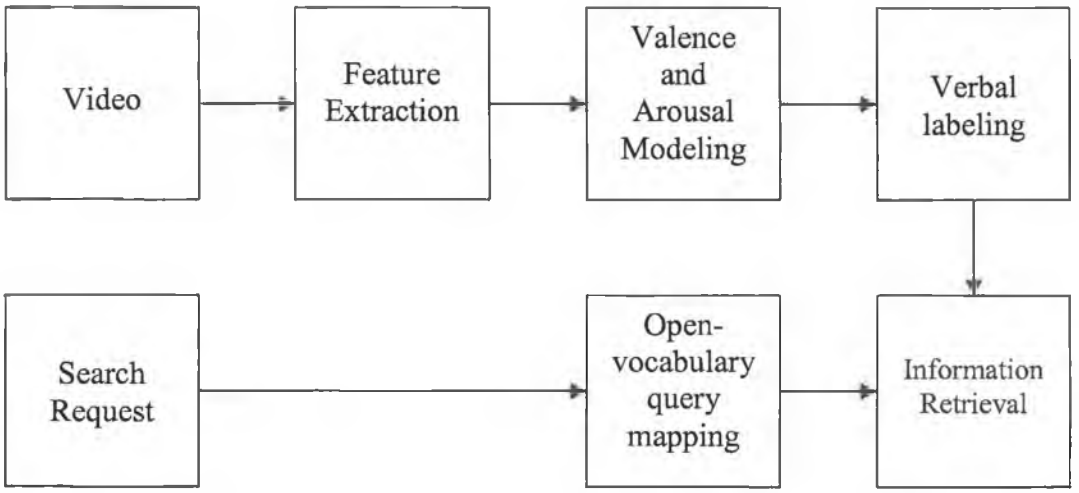


Figure 6-1. System Overview.

As shown in Figure 6-1, the system is divided into 5 components; feature extraction, valence and arousal modeling, verbal labeling, IR and open-vocabulary query mapping. Implementation of each of the 5 components will be described in the following sections.

The system processes the videos to extract low-level visual and audio features. These visual and audio features are passed on to the next component where they are combined and used to model valence and arousal. From there on, using valence and arousal, verbal emotional labels are detected and indexed for the videos. These labels are then stored and used whenever a user makes a query. The user inputs a text query which goes through the open-vocabulary query mapping component to map the query words onto the system's set of pre-defined emotional labels which are then passed on to the IR component of the system. The IR system then processes the query and returns a list of sorted results that corresponds to the video clips that most matches the query.

6.4 Feature Extraction

The feature extraction component is the first step of the processing. This component extracts the low-level features that are required for the valence and arousal modeling in the next step.

For visual features, each frame of the visual stream was decompressed and processed using Matlab's built-in function "aviread". This function reads in each frame of the video and converts it into a form that can be manipulated within Matlab. The .wav uncompressed audio file is read using Matlab's built-in function "wavread". Following the norms of audio analysis, the audio values are read and processed in frames of 20 ms in duration and with 2/3 overlap between adjacent frames. This is to ensure that relatively short changes even between frames are captured. Once these visual and audio values are obtained, they can be processed to generate low-level features.

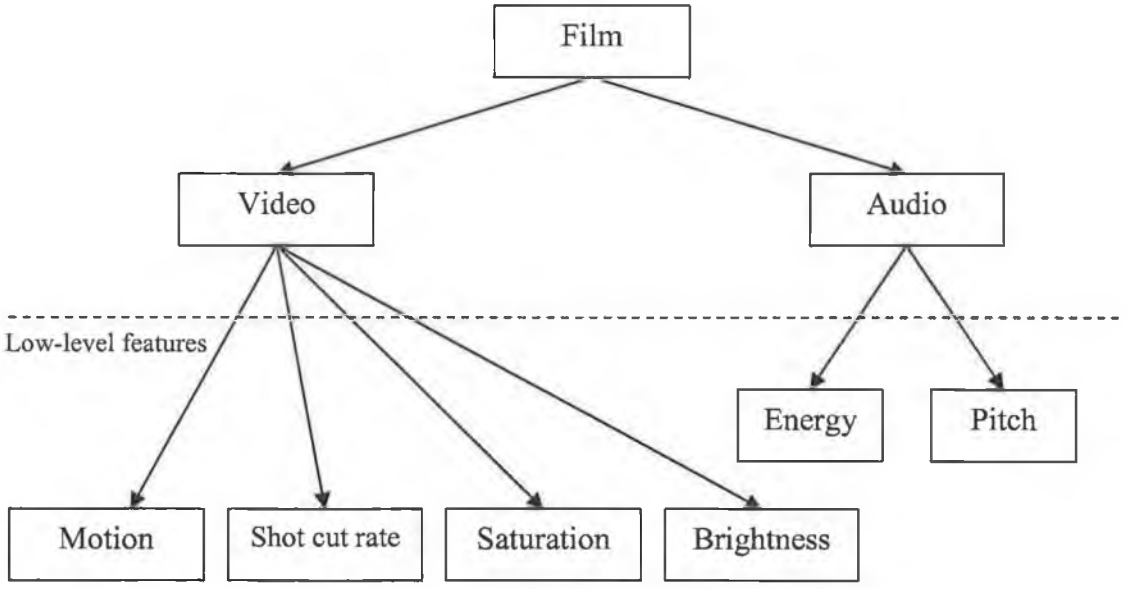


Figure 6-2. Low-level feature extraction from video and audio of a film

As shown in Figure 6-2, 6 low-level features are extracted from the audio and visual features. For the visual stream, 4 features are extracted: motion, shot cuts, saturation, and brightness. For audio, only the audio energy and pitch is extracted. The

justifications for using these six features were explained in Section 4.8 (Related Work in Detecting Emotions from Films). In Hanjalic and Xu [2003], modeling of arousal was based on three features (motion, shot cut rate and energy), using only one audio feature (pitch) for modeling valence seems unbalanced and leaves room for improvement. It seems intuitive to assume that visual cues in a piece of video would be a key influence in the viewer's feelings or valence.

Therefore this thesis proposes to incorporate preliminary work and findings by [Kok, 2006] to use colour saturation and brightness as extra visual features to model valence. This aims to improve the accuracy and reliability of the valence modeling as it was found during the development of the affect-based retrieval system that valence based only on pitch is susceptible to large fluctuations which might not correspond with what is going on in the video. For example, in scenes with high-pitched human shouting or sound effects, the dark colours of the scene could indicate a horror or terror scene, while bright colours could indicate a funny and happy scene.

The selection of these features is made in line with cinematographic guidelines and intuition in order to ensure that each feature contributes purposefully to valence and arousal modeling. However the subjective nature of evaluating the performance of valence and arousal modeling makes it difficult to determine the value of the respective features during development of the system since there are no previous results available relating to similar systems. Therefore a system needs to be built and a suitable retrieval test set established first before the respective features can be tested for their importance in future work.

- **Motion**

The motion feature calculated here is the overall motion activity computed between consecutive video frames. The computed values, called motion vectors, are calculated using a standard block-matching motion estimation method between two frames of k and k+1. As described in Section 2.4.5 (Motion Estimation and Compensation), the aim of motion estimation is to search for a macroblock in a search video frame that matches the macroblock in the reference video frame. The calculated displacement values are then

called the motion vectors. For this system, only the magnitude of the motion is significant, so the direction of the motion vectors is not important and is not used.

While MPEG-encoded video files could have provided these motion vectors [Gilvarry, 1999], it was decided that a full search block matching algorithm should be implemented in Matlab for learning and development purposes. Many other options are available for deriving the motion feature, however for this system's purpose the difference in using other methods are not expected to be significant. The full search algorithm was later replaced by a downloaded Matlab script file containing some block-matching algorithms that work similarly, but trades off a little accuracy for speed when a large number of videos are to be processed. The downloaded algorithms included full search, three-step search, new three-step search, simple and efficient search, four step search, diamond search, and adaptive rood pattern search algorithms.

These algorithms are widely accepted by the video compressing community and have been used in various standards, ranging from MPEG1 to MPEG4. The diamond search algorithm was used for this system as it has been observed to be the best block matching algorithm and all newer search algorithms are improvements based on the original diamond search algorithm [Barjatya, 2004].

Therefore for each frame of video, the overall average motion is computed between two adjacent frames k and k+1 using the diamond search algorithm. The motion activity value m(k) is then found as the average magnitude of all (B in total) motion vectors $v_i(k)$, scaled to the range between 0 and 100% and then normalized by the maximum possible length of a motion vector $|v_{max}|$ to derive the scaled average motion function m'(k).

$$m'(k) = \frac{100}{B \times |v_{max}|} \times \sum_{i=1}^{B} |v_i(k)| \quad \%$$

Figure 6-3 shows a plot of m'(k),

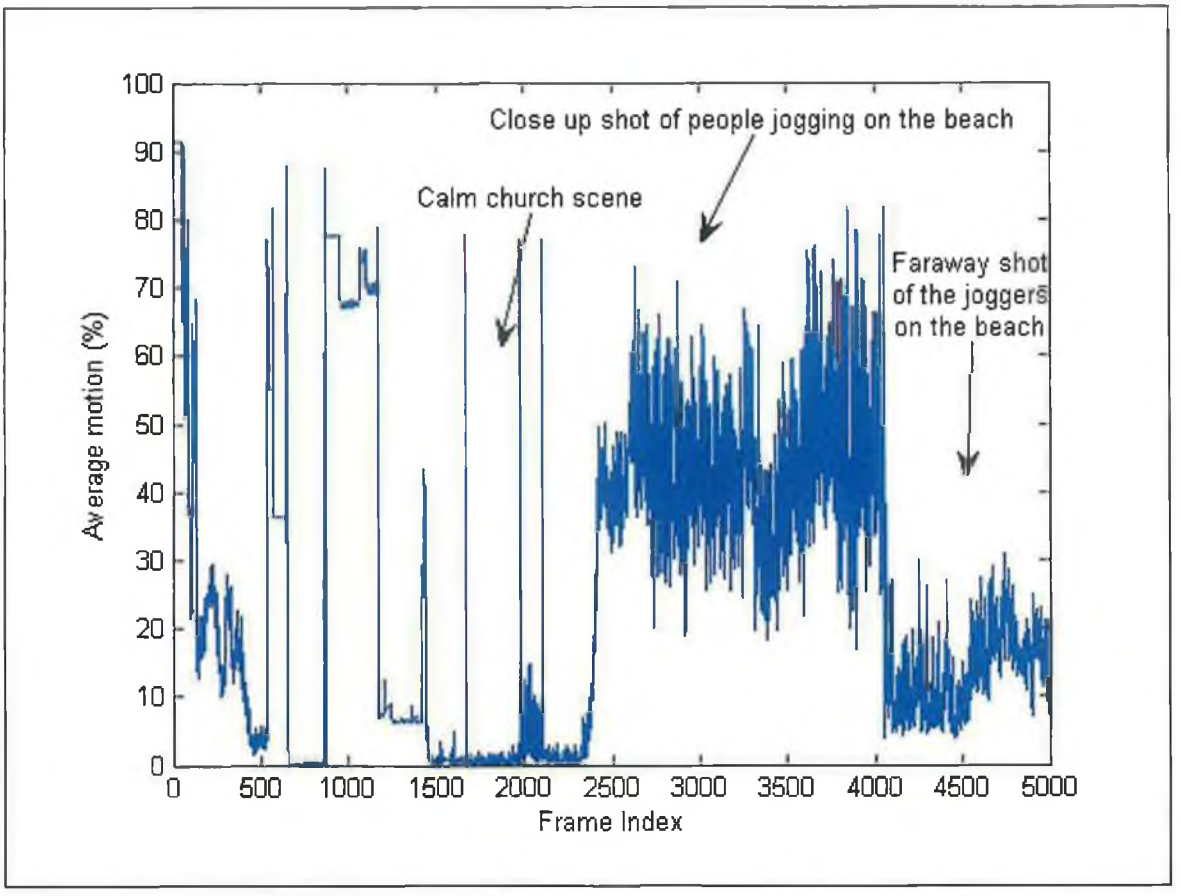


Figure 6-3. Plot of average motion m'(k) for the first few minutes of the movie "Chariots of Fire"

Figure 6-3 shows an average motion plot showing how much activity there is in a scene. The plot above is taken from the first three and a half minutes of the movie "Chariots of Fire". From frames 1 to 1500, the scenes are of various credits and animation for the movie companies. This is followed by calm, mostly static scenes of a person addressing a group of people in a church until about frame 2500. The spikes in-between are due to the shot cuts, which are detected as an abrupt change in motion. However, these solitary spikes are less pronounced when a smoothing and averaging function is applied at a later stage. From frame 2500 onwards, the scene changes to a close up shot of a group of people jogging on the beach. This scene is dynamic as the camera bounces around while tracking the faces of the individuals as they jog. Therefore the average motion is quite sustained at a high rate until the frame 4000. From then on the

scene changes to a more faraway shot of joggers where the camera is stable and not moving. This is reflected in the plot as the average motion drops to a lower but still sustained level.

- **Shot Cut Rate**

While it is called shot cut rate, an alternative name for this feature would be "rhythm". This is because similar to the analysis of motion activity, the aim is to obtain a curve that is a function of the frame index k that reveals a connection between a viewer's arousal and the time-varying shot lengths.

The first step is to actually detect the shot cuts that occur in the video. A shot cut is detected using a colour histogram approach that has been shown to perform reliably by [Browne et al., 2002]. Similar to motion, adjacent frames are compared and a shot cut is detected when the difference between the colours in the two frames exceeds a pre-determined threshold. More sophisticated shot boundary detection methods exist, but here a simple and fairly robust method was used. A large amount of data was to be processed and small numbers of errors can be tolerated since only a measure of shot cut rate is required.

As in Section 2.1.4 (Colour Space Representation in Digital Video), each video frame arrives in RGB format, and is converted into the YUV colour space. For each component Y, U and V, a 64-bin colour histogram is computed and then concatenated to form a 192 unit vector. This process is repeated for the video frame that is to be compared against. The end result is a pair of vectors that represent the colour information of each frame.

The comparison is then done using the dissimilarity analogue of the cosine similarity measure (CSM), given two histograms A and B,

$$D_{cos}(A,B) = 1 - \frac{\sum_{i=1}^{N} (a_i \times b_i)}{\left(\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2\right)}$$

The output $D_{cos}(A,B)$ gives the cosine of the angle between the two vectors. This can also be interpreted as the difference between the two frame's histogram. Figure 6-4 shows A plot of the histogram difference. By setting a threshold (represented by the red line), a shot cut is said to occur when the D_{cos} value exceeds it [O'Toole et al., 1999].

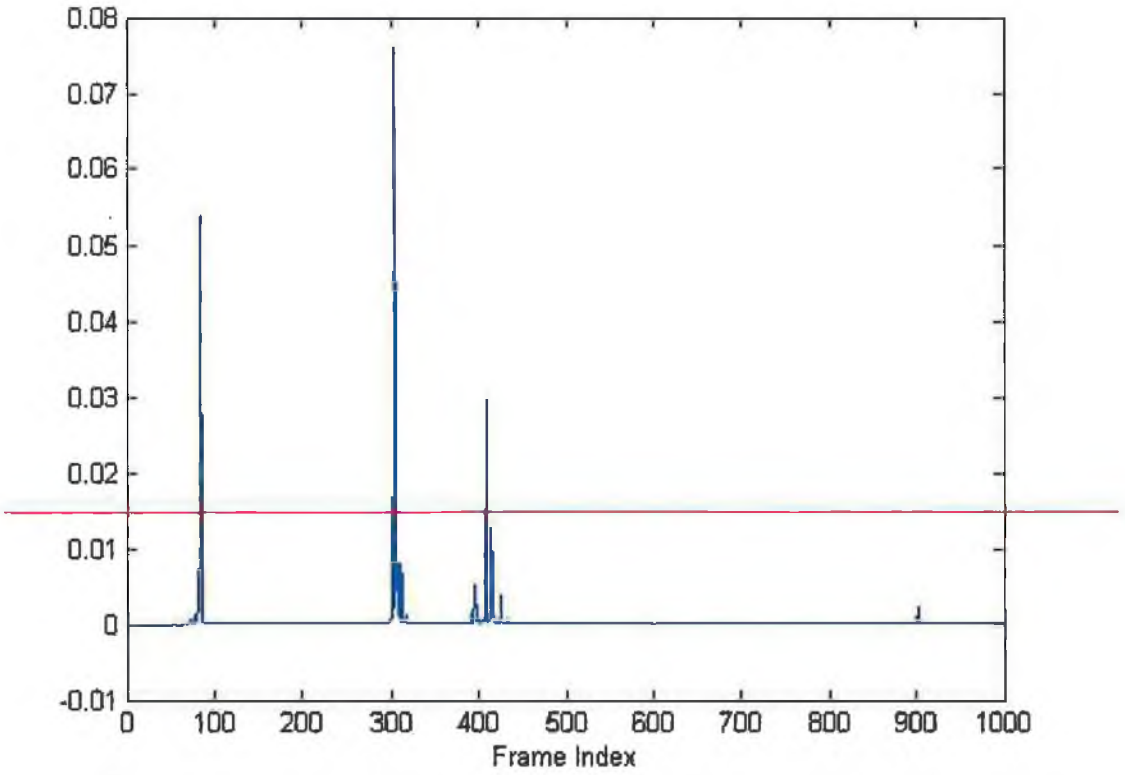


Figure 6-4. A plot of the cosine similarity values or histogram difference. The red line is a manually set threshold to detect shot cuts, which detected 3 shot cuts.

After the shot cuts have been detected along the video, the shot cut rate is modeled by defining the scaled shot cut rate function c'(k),

$$c'(k) = 100 \times e^{(1-(n(k)-p(k)))}\ \%$$

Here, p(k) and n(k) are the positions (frame indexes) of the two closest shot cuts to the left and right of the frame k, respectively, with c'(k) values distributed on the scale between 0% to 100%. The result is a step curve, with each step corresponding to a video segment between two shot cuts and the height of each step being inversely related to the interval between the boundaries, therefore the shorter the interval, the higher the value c'(k).

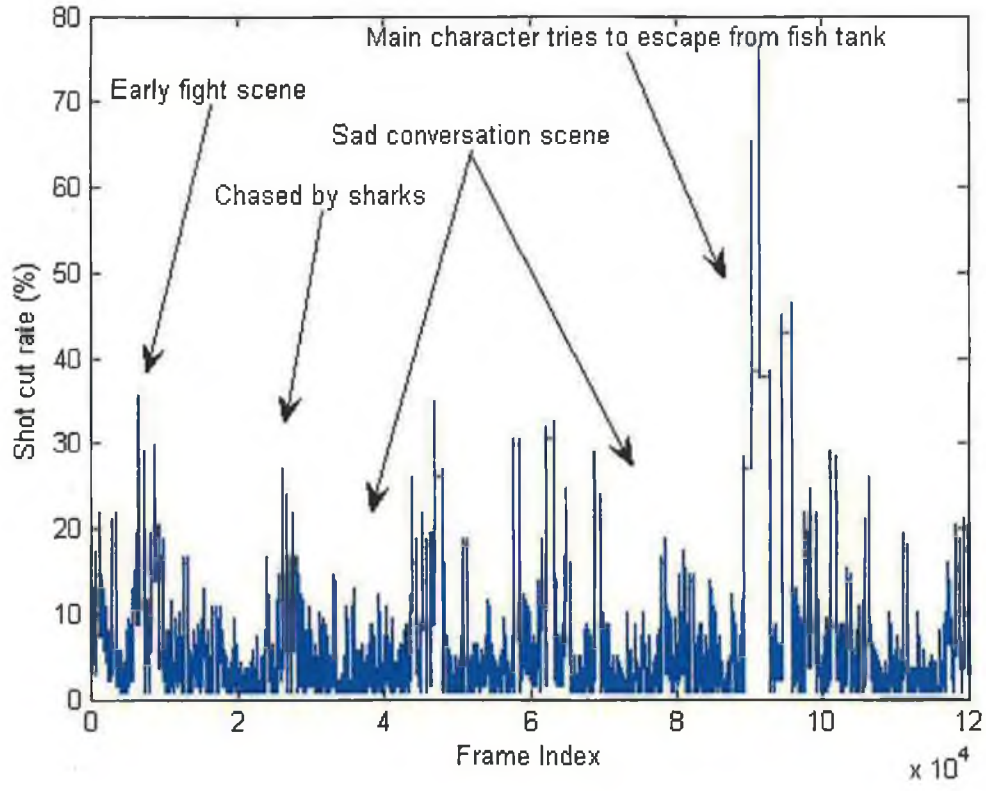


Figure 6-5. A step curve c'(k) of the movie "Finding Nemo".

The example shown in Figure 6-5 illustrates the shot cut rate indicating scenes of high activity. This is because a director of a film often uses a rapid succession of shots not only to portray as much information about the action in the short amount of time that it occurs for, but also to add suspense and grab viewer attention. The plot in Figure 6-5 shows the shot cut rate for the movie "Finding Nemo". It can be observed that at areas where there is a concentration of higher peaks of shot cuts, it usually points to the more action-oriented or high activity scenes. In contrast, conversation scenes reveal a more stable and lower rate of sustained shot cuts.

- **Saturation**

The HSV colour space is derived from the RGB colour space, and can be useful in revealing information about how vibrant a colour is. In order to obtain a curve s(k) for saturation, each frame of the video is converted into the HSV colour space. Then the values (saturation) are summed and the average value in the colour space is taken and scaled between 0% to 100% to derive the scaled average saturation function s'(k).

$$s'(k) = 100 \times \frac{s(k)}{(s_{max} - s_{min})} \%$$

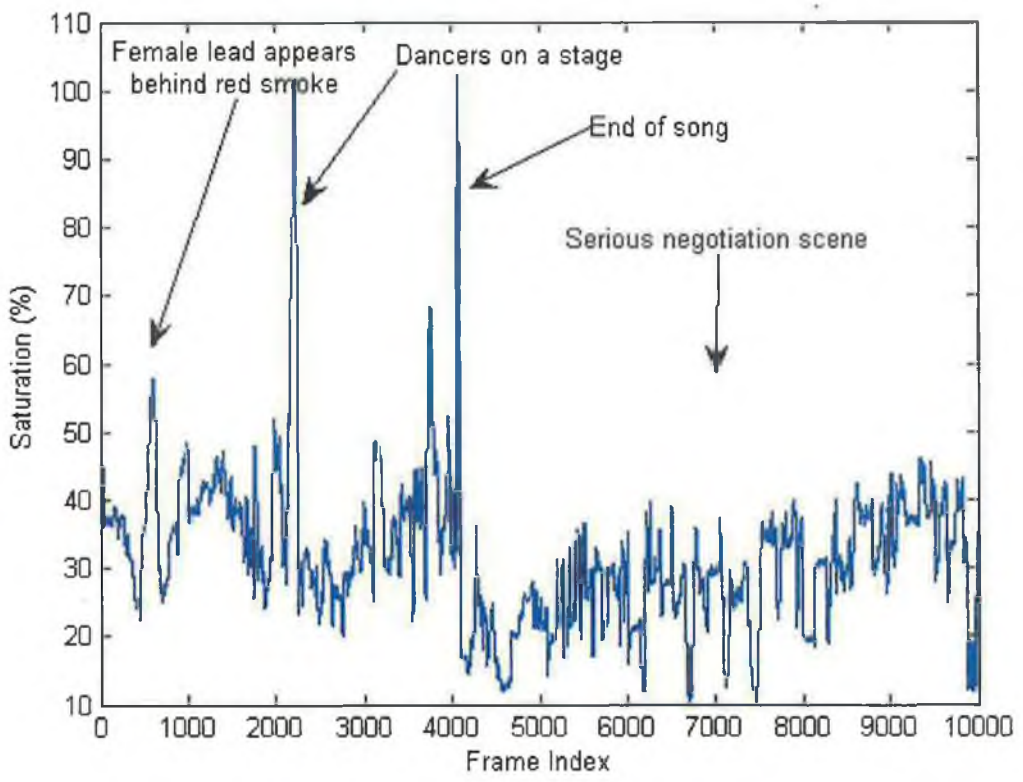


Figure 6-6. A plot of the saturation curve, s'(k) for "Indiana Jones: Temple of Doom".

The curve in Figure 6-6 plots the first six and a half minutes of the film "Indiana Jones and the Temple of Doom". The saturation curve reveals how vibrant the colours are, as can be seen in the plot above. This section of the film is a song and dance scene followed by a tense negotiation scene. The first scene opens with the female lead character appearing from behind a thick cloud of red smoke hence a noticeable spike. Later on, the scene shows dancers wearing bright flashy jackets, in the middle as can be seen as a bigger peak in the plot. The peak near the end of the song corresponds to a brilliant flash of red that signifies the end of the song, after which the scene shifts to a more serious negotiation scene with muted colours. The lack of any prominent peaks as well as a slightly lower average values compared to the first scene of the plots shows that the more serious scene is more muted in terms of colour vibrancy.

- **Brightness**

Similar to saturation, the brightness feature is derived from the V values of the HSV colour space to represent how bright a colour is. Therefore, to obtain the curve b(k) for brightness, each frame of the video is converted into the HSV colour space. The values of b(k) are then summed, and the average value in the colour space is taken and scaled between 0% to 100% to derive the scaled average brightness function b'(k).

$$b'(k) = 100 \times \frac{b(k)}{(b_{max} - b_{min})} \%$$

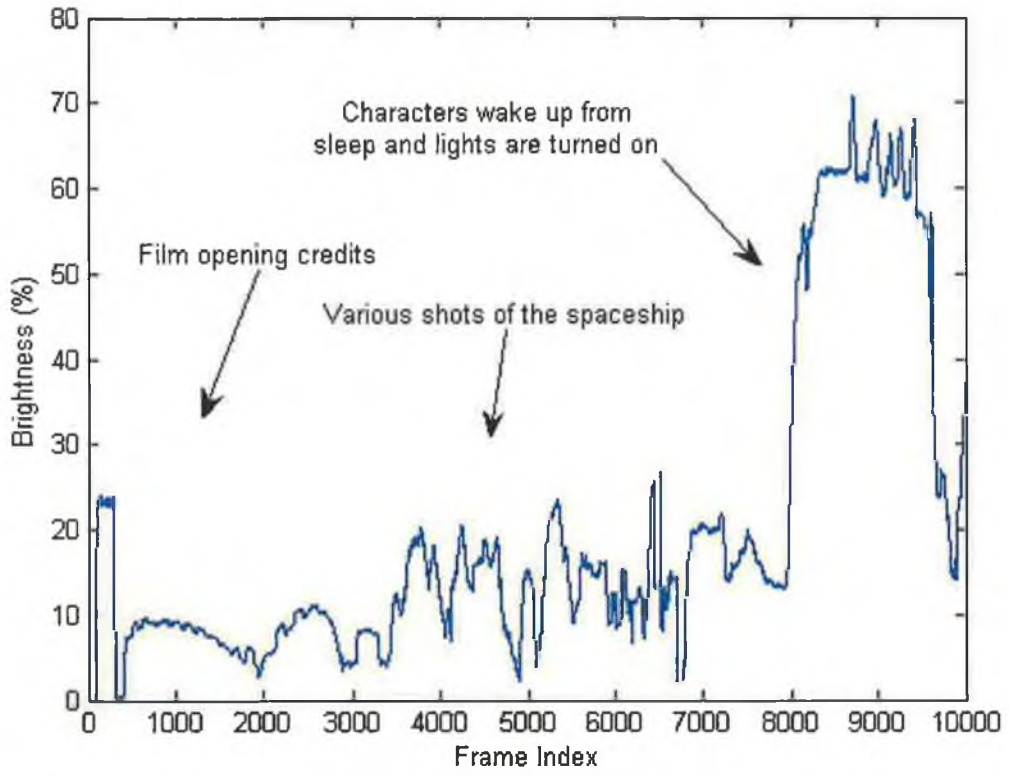


Figure 6-7. A plot of the brightness curve b'(k) for first few minutes of "Alien".

An example of a brightness curve is shown in Figure 6-7. This shows the first 6 and a half minutes of the film "Alien". The majority of the scene is rather dark, where the film opening credits have a black background with overlaid text appearing occasionally. This is then followed by various shots of a spaceship drifting through space, which still retains a dark atmosphere until around frame 8000, at this point the main characters awake from their space sleep and the lights are turned on, producing a bright scene.

- **Energy**

For audio features one value is computed for each frame of audio that is processed. Therefore each frame will yield one value of the short time energy function e(k). Where w(m) is a rectangle window, e(k) is defined as,

$$e(k) = \frac{1}{N} \times \sum_{i=1}^{N} (x(m) \times w(n-m))^2$$

This is then scaled between 0% to 100% to derive the scaled short time energy function e'(k),

$$e'(k) = 100 \times \frac{e(k)}{(e_{max} - e_{min})} \%$$

However, because of the size and overlapping of frames, audio features will have more than double the amount of values compared to visual features. In order to make sure both features are synchronized, the audio values are re-sampled into the same total number of values as the visual features.

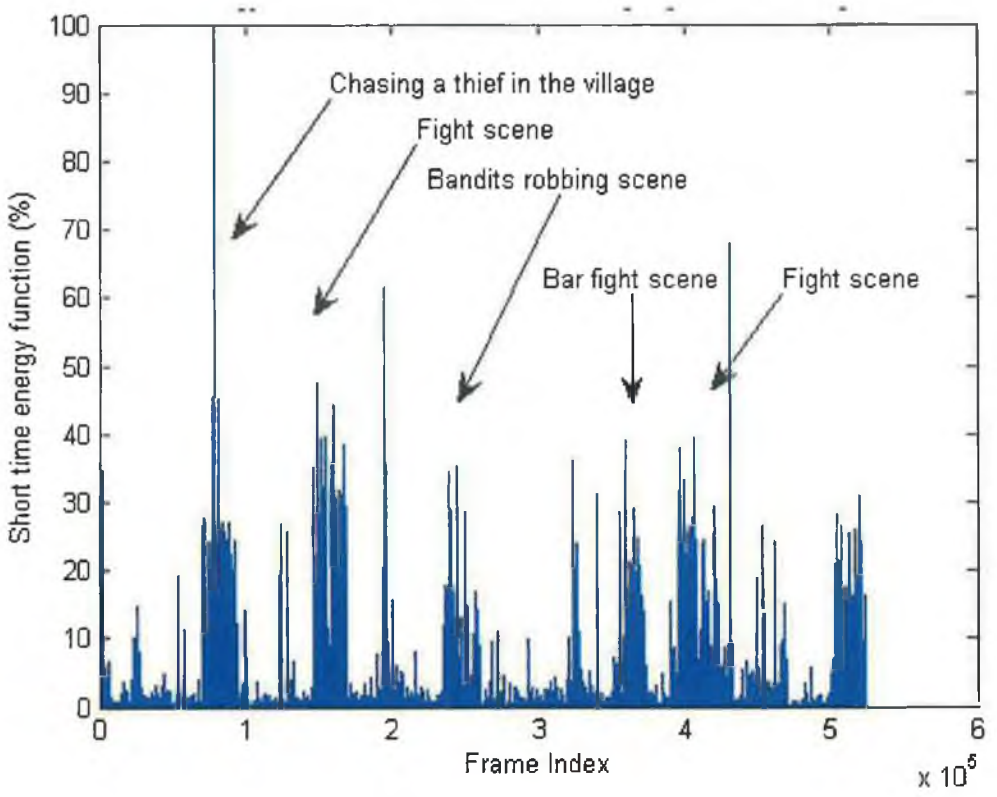


Figure 6-8. A plot of the energy curve e'(k) for the movie "Crouching Tiger Hidden Dragon".

Figure 7-8 shows an example of an energy curve. In this plot of the entire kung-fu film "Crouching Tiger Hidden Dragon", almost all of the scenes that involve fighting show higher concentrations of energy, as can be seen in the graph. Parts of the film which do not involve fighting exhibit lower audio energy values, therefore suggesting a much lower level of activity. These lower level of audio energy segments usually correspond to scenes of conversations between characters and other scenes where there is plot progression but little on-screen activity. This is explained by the deliberate use of louder music and sound effects during scenes or interest or with high levels of activity.

- **Pitch**

Pitch tracking is a difficult task that can be approached using a wide range of techniques. In their work, Hanjalic and Xu, [2003] used pitch only for voiced segments of video, the

approach adopted in this thesis calculates the pitch regardless of any voiced or unvoiced segments. In order to be able to label segments of video that do not contain any voice or speech, the approach here is to assume that the original assumption made by Hanjalic and Xu that higher pitch in vocal segments equals higher valence still holds true when applied to non-vocal segments.

Pitch detection or its equivalent, the fundamental frequency estimation, is one of the most important problems in both speech signal and music content analysis. Although equivalent, pitch detection is a rather more perceptual term, while fundamental frequency is a physical measurement. Furthermore, the pitch tracking methods is not explained in detail by [Hanjalic and Xu, 2003], other than a mention of off-the-shelve pitch-tracker software. Because the pitch will be derived from all sections of audio and there can be a wide range of audio events in videos, the pitch feature here follows the approach taken in [Zhang and Jay Kuo, 2001], which according to the authors is "to build a method which is efficient, robust, but not necessarily perfectly precise" that works satisfactorily for a wide range of audio signals.

Therefore the pitch or fundamental frequency is calculated based on peak detection from the spectrum of the sound. The power spectrum is generated with autoregressive (AR) model coefficients estimated from the autocorrelation of audio signals $R_n(k)$. Thus the autocorrelation of the audio signal can be computed as,

$$R_n(k) = \frac{1}{N} \times \sum_{i=1}^{N} x(m)) \times (x(m+k)) \text{ ,where } 0<k<P$$

where P is the order of the AR model. The AR model parameters are estimated from the values of $R_n(k)$ through the Levinson-Durbin algorithm [Haykin, 1991]. Then the power spectrum S_{AR} can be estimated as,

$$S_{AR}(e^{j(2\pi/N)l}) = \frac{E_P}{\left| \sum_{k=0}^{N-1} a_k \times e^{j(2\pi/N)l} \right|^2}$$

where E_P is the mean-square prediction error, and a_k the prediction coefficients. The denominator of S_{AR} can be computed with an N-point FFT. N is set to 512 to ensure reasonable frequency resolution. This AR model generated spectrum is a smoothed version of the frequency representation. Moreover, the AR model is an all-pole expression, making the peaks prominent in the spectrum. The order (P) of the AR model was chosen to be 40 as suggested by [Zhang and Jay Kuo, 2001] to obtain good precision of the fundamental frequency.

The highest peak in the spectrum is chosen as the fundamental frequency for each frame to obtain a curve p(k) and then plotted on a graph to obtain a pitch track, which is then scaled again onto the range 0% to 100% to derive the scaled fundamental frequency function p'(k),

$$p'(k) = 100 \times \frac{p(k)}{(p_{max} - p_{min})} \%$$

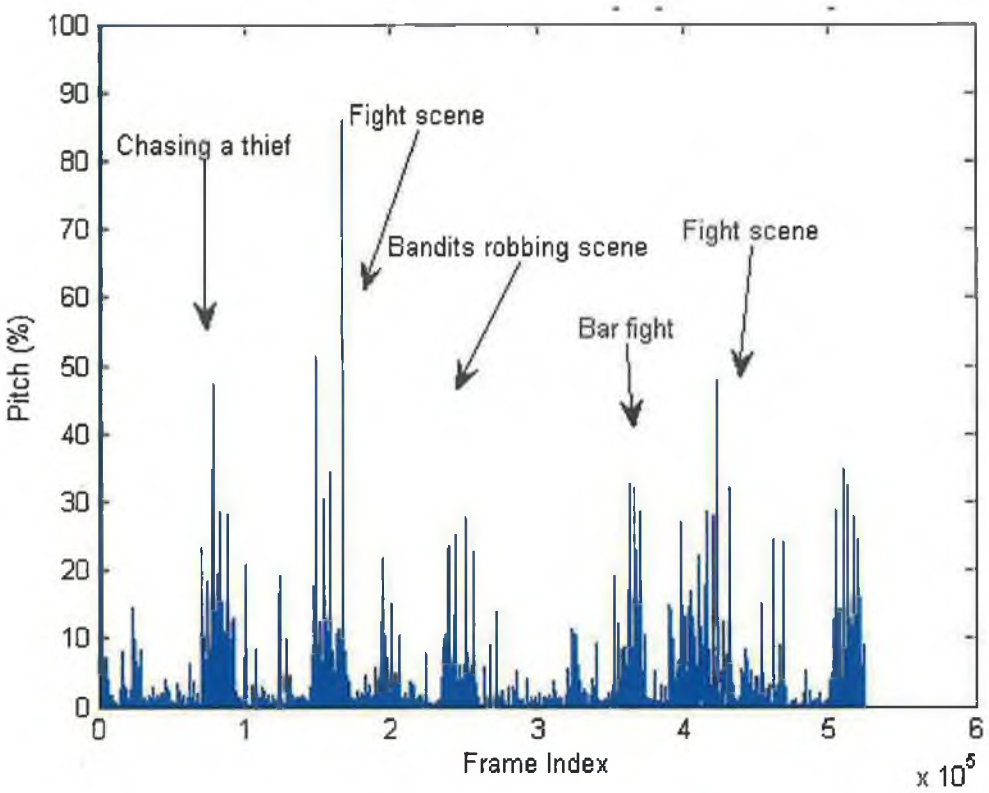


Figure 6-9. A plot of the pitch curve p'(k) for the movie "Crouching Tiger Hidden Dragon".

Figure 6-9 shows an example of a pitch curve which tracks the frequencies occurring in the video. Higher frequencies will then show up as higher values. The curve shows the pitch values for the entire film "Crouching Tiger Hidden Dragon". Due to the occurrence of pronounced music and sound effects with higher pitches, especially for scenes with a high level of activity in this film, the pitch also shows higher values for the same scenes where energy was high.

6.5 The Three Criteria for Modeling Valence and Arousal

Hanjalic and Xu [2003] established three criteria that should be considered when modeling valence and arousal. The criteria are:

- **Comparability**

The criterion "comparability" ensures that the values of the arousal and valence obtained in different videos are comparable. This means that all values have to be normalized and scaled appropriately so they are directly comparable. This criterion is fulfilled as all features were normalized and scaled onto the range 0% to 100%.

- **Compatibility**

The second criterion "compatibility" ensures that the curves cover an area in the affect curve which corresponds to the VA (Valence and Arousal) emotion space. While [Hanjalic and Xu, 2003] illustrates that the shape and distribution of emotional labels on the VA emotion space have a parabolic-like contour, in the affect-based retrieval system presented in this thesis, the emotional labels that were used and their distribution (between 0 and 1) were fairly distributed across the entire on the VA emotion space (Section 6.7 Verbal Labeling). Therefore the arousal and valence values that are calculated here can be linearly mapped to the emotion space between 0 and 1 and still fulfill the criterion.

- **Smoothness**

The third criterion "smoothness" accounts for the degree of memory retention of preceding frames and shots. This means that the perception of the content does not change abruptly from one video frame to another, but is a function of a number of consecutive frames. For example, this means that although low-level features can fluctuate wildly between frames, human emotion perception does not follow such rapid changes, as there is a certain "inertia" to the change in affective interpretation. Therefore a smoothing function has to be applied to the features to fulfill this criterion.

All low-level features therefore have to undergo a smoothing function by convolving the features with a suitably long Kaiser window $K(l_1,\beta_1)$ as shown in Figure 6-10. The l_1 is the length of the window and the β_1 value is the shape parameter. By

subjective evaluation, it is found that setting the l_1 of the Kaiser window to be 750, or approximately 30 second duration window with a shape parameter $\beta_1 = 5$ yielded the best result, which is comparable to the result in Hanjalic and Xu's [2003] result.

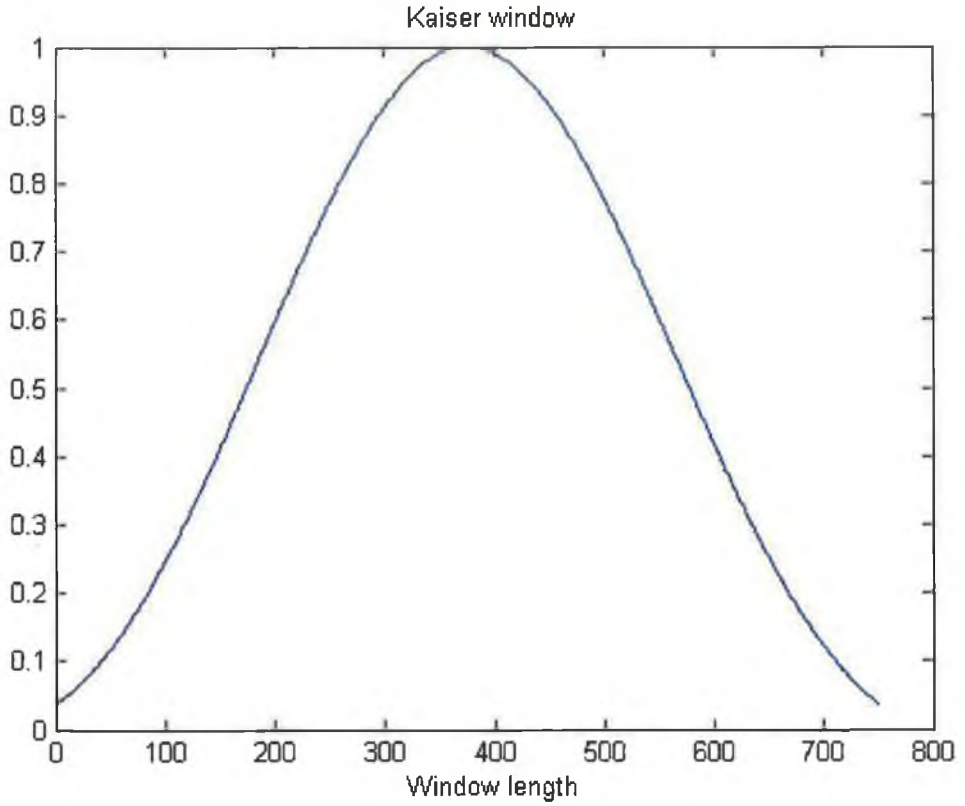


Figure 6-10. A Kaiser window with $l_1 = 750$ and $\beta_1 = 5$.

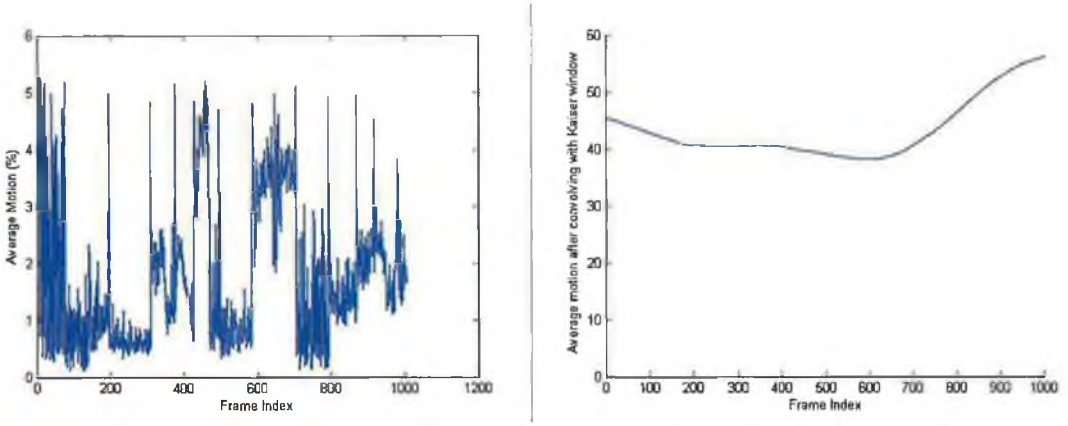


Figure 6-11. Motion feature before (left) and after (right) convolving with a Kaiser window

Figure 6-11 shows the motion feature before and after it has been convolved with a Kaiser window. The wild fluctuations are smoothed to produce a stable line that shows the underlying trend of the motion feature.

6.6 Valence and Arousal Modeling

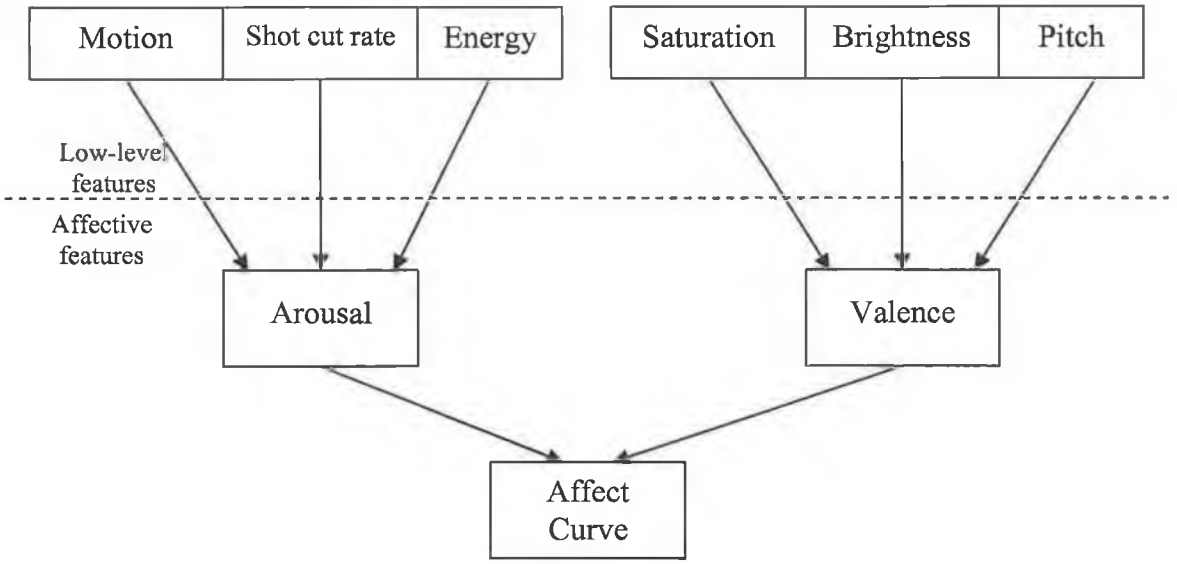


Figure 6-12. Low-level features used to model valence and arousal and affect curve.

After smoothing, the features are ready to be combined to form valence and arousal curves as shown in Figure 6-12. Arousal A(k) is defined as the weighted linear sum of M(k), E(k) and C(k),

$$A(k) = \frac{(\alpha \times M(k)) + (\beta \times E(k)) + (\gamma \times C(k))}{(\alpha + \beta + \gamma)}$$

where M(k), E(k), and C(k) are the smoothed forms of m(k), e(k) and c(k) respectively and α, β, and γ are the weighting factors. Through subjective evaluations involving random video clips from several movies, it was found that setting all weights to 1 yielded the best practical result. This can be justified due to the wide range of films that are being analyzed in this study and the results of human feedback on the performance of the arousal curves in the experimental section (Section 7.6 Results and Discussion of the User Evaluation).

Valence is similarly defined as the weighted sum of B(k), S(k) and P(k),

$$V(k) = \frac{(\alpha \times B(k)) + (\beta \times S(k)) + (\gamma \times P(k))}{(\alpha + \beta + \gamma)}$$

where B(k), S(k), and P(k) are the smoothed forms of b(k), s(k) and p(k) respectively and α, β, and γ are the weighting factors. The weights are set as $\alpha = 0.69$, $\beta = 0.22$ and $\gamma = 1$ respectively. The different weights for saturation and brightness follow the findings of [Kok, 2006], where the values were derived from Valdez and Mehrabian's [1994] work on how saturation and brightness were related to the VA emotion space. Through subjective evaluations again, it was determined that these weights worked well for a wide range of films.

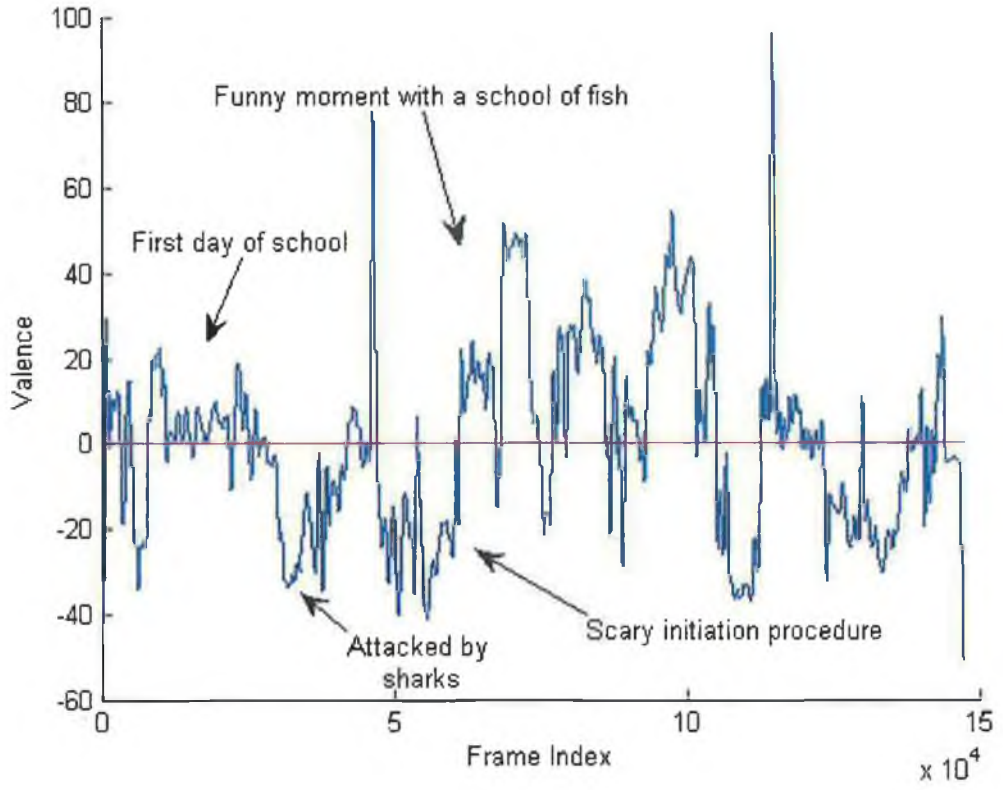


Figure 6-13. Some description of the valence curve for the movie "Finding Nemo".

Figure 6-13 shows an example of a valence curve for the entire film "Finding Nemo". The "Finding Nemo" film is a comedy film featuring fishes as the main characters. The red line is the "neutral" value. This means that the closer the curves are to the red line, the more neutral the scene is in terms of valence. The higher the curve, the more positive in terms of valence the video is.

In the valence curve, it can be observed that during the beginning of the movie, the main characters were in an uplifting mood as they prepare for the first day of school; therefore the curve is in the positive feeling region. When the main character is attacked and chased by sharks in the dark murky waters, the curve drops below neutral and into the negative feeling regions. When the main character undergoes a scary initiation procedure, the curve drops into the negative feeling region again. However, the curve rises up to the positive feeling region again when the main character encounters some fish and experiences a funny moment. Through such subjective evaluation and reasoning, it

can be said that the valence curve is working reasonably well in describing the emotional content of the film.

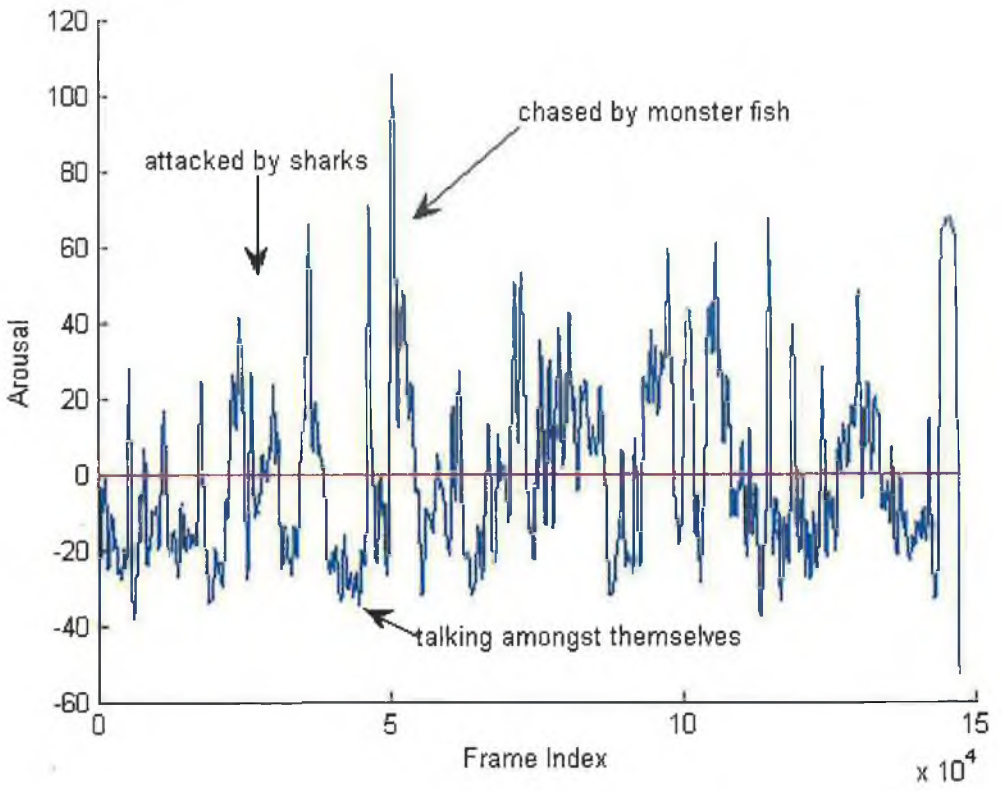


Figure 6-14. Some description of the arousal curve of the movie "Finding Nemo"

Figure 6-14 shows the corresponding arousal curve, while it seems that there are a lot of peaks and fluctuations, the curve is actually quite compressed as it plots the entire movie. When zoomed in, the lines are smooth (as a result of the convolution) and seem to correlate well with the movie. For example, during the scene where the character is chased by sharks, there is high and sustained activity in the arousal curve. When the characters are talking amongst themselves with little activity, the curve drops to the negative region, suggesting a more relaxed atmosphere. When the main character is suddenly frightened by a scary monster fish, a sudden high peak is observed after some period of low activity.

The red line was derived from the average of the values of the whole curve, therefore making the assumption that in a suitably long piece of video such as a film, the total emotional content varies equally from negative to positive. While different genres of films might have different bias towards different emotions, this simplifying assumption is taken to facilitate the almost completely automatic nature of this system. The subjective evaluations revealed that this approach works suitably well for the wide range of videos, especially for film data, and is further justified in the human feedback on arousal and valence curves in Section 7.6 (Results and Discussion of the User Evaluation).

The valence and arousal curves can be combined to form the affect curve and mapped onto the VA emotion space as shown in Figure 6-15. Using the x-axis for valence and y-axis for arousal, the curve starts off at a point and then travels through the VA emotion space "snaking" between different regions of the VA emotion space until it comes to a halt at the end of the film. Therefore each point of the affect curve occupies a point in the VA emotion space which can then be associated with the emotional labels that happen to be nearest to this point on the affect curve.

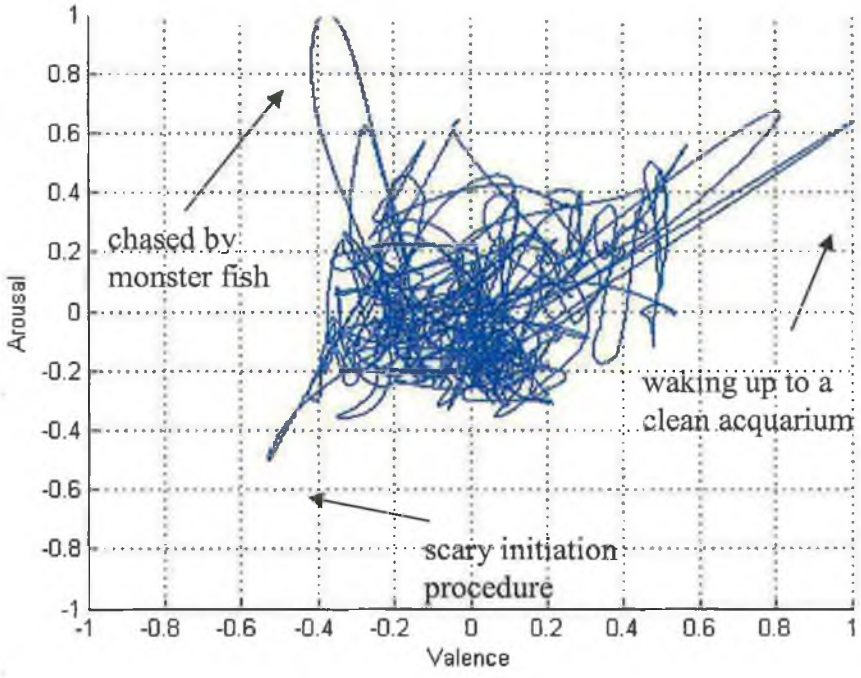


Figure 6-15. Affect curve for the movie "Finding Nemo" on the VA emotion space.

Figure 6-15 shows how the line occupies different regions of the affect curve for the different types of scenes from the movie. For example, during the scene where the main characters were chased by a monster fish, the low valence but high arousal curves (Figure 6-13 and 6-14) plots the affect curve in the top left region of the VA emotion space, which is typically associated with the emotion "Fear" and "Angry" as seen in Figure 4-3 in Section 4.2 (VAD Emotion Space).

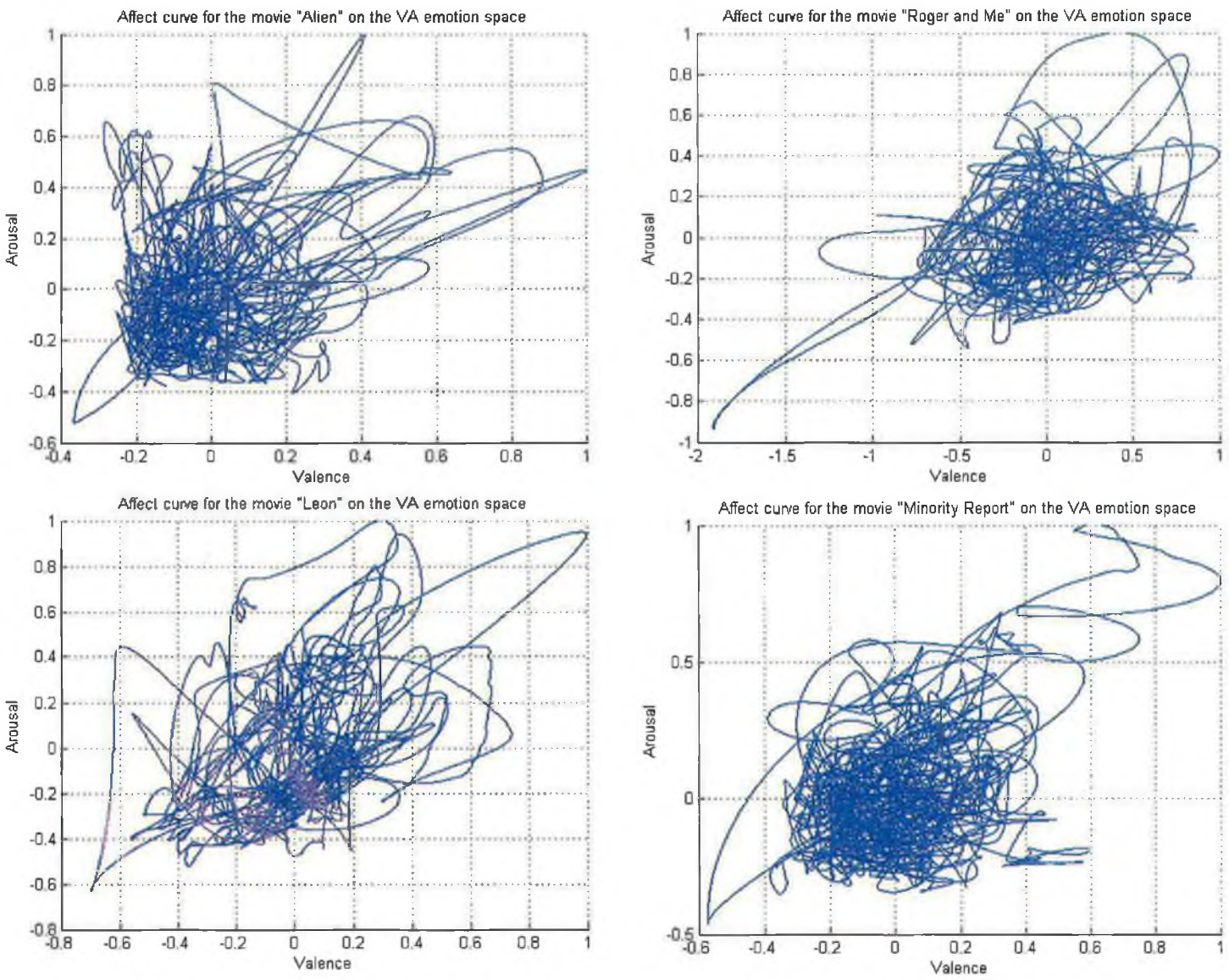


Figure 6-16. Affect curve for (a) "Alien", (b) "Roger and Me", (c) "Leon", and (d) "Minority Report" on the VA emotion space.

Figure 6-16 shows affect curves for the movies "Alien", "Roger and Me", "Leon", and "Minority Report". As can be observed from Figure 6-16, the shape of the affect curve is different for each film, therefore revealing the emotional development of the films. For example, the affect curve for the horror film "Alien" is mainly concentrated in the negative regions of valence and arousal, while other films have a more equal spread in the VA emotion space. In addition, the peaks of the affect curve show that the film "Leon" contains a wider range of dominant emotions as the affect curve is less concentrated on a small area in the VA emotion space. For future work, the shape and the location of these peaks could be used to locate interesting scenes or highlights of the films.

6.7 Verbal Labeling

In order to automatically index video with labels that describe the affective states of the video, the VA emotion space has to be populated with emotional verbal labels. Related work concerning populating the VA emotion space with emotional labels is lacking. Related work such as Hang's [2003] that uses the VA emotion space uses a small number of very broad terms for classification of multimedia data like "happy", "sad", and "fear".

In an attempt to further extend the usability of the VA emotion space, this thesis therefore proposes a novel use of the findings of a study by Russell and Mehrabian [1977]. This populates the VA emotion space with a far greater number of labels that permit a finer granularity of emotion analysis. Their study was conducted with human subjects where they were asked to rate 151 words along the dimension of valence and arousal. For each word, they were required to rate on a scale where the word should fall on the valence and arousal dimension. These values are then normalized between the range -1 and +1. Using the values from their findings, the VA emotion space is populated with the 151 words or verbal labels. Some example of the labels are "bold", "inspired", "quiet", "shy" and "sad". These labels were chosen to try to represent the entire wide range of emotions, therefore the 151 words is quite an extensive list of labels. Each label comes with arousal and valence values, so "bold" would have arousal value of 0.61 and valence value of 0.44, and "nonchalant" would have arousal value of -0.25 and valence of

0.07. Figure 6-16 shows the location of these two labels on the VA emotion space. This list of labels is referred to as the "Russell List".

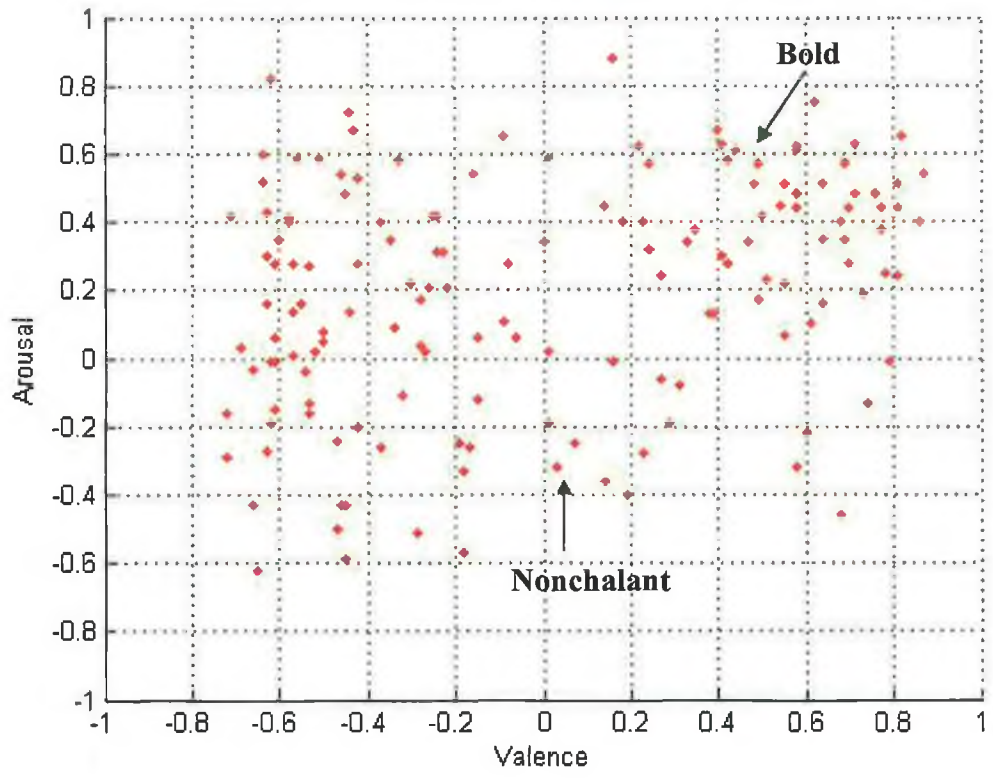


Figure 6-17. The 151 emotional verbal labels plotted as red dots in the VA emotion space.

As can be seen in Figure 6-17 on the VA emotion space, the values are quite fairly distributed, therefore eliminating the need to normalize the affect curve other than scaling it to between the same range of -1 and +1 as well as fulfilling the compatibility criterion.

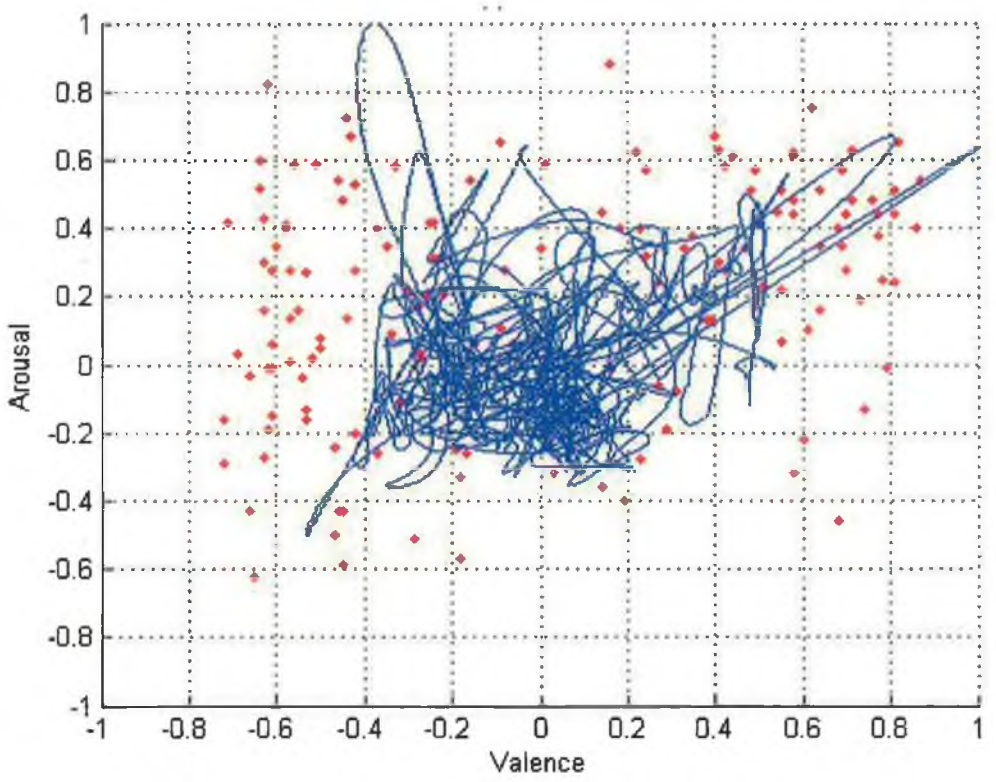


Figure 6-18. Plotting the affect curve and the 151 words.

As can be seen in Figure 6-18, each point on the affect curve can then be associated with the closest word or verbal label from Russell and Mehrabian's [1977] study. The end result of this mapping is a list of detected verbal labels with one assigned to each frame of the film from start to end.

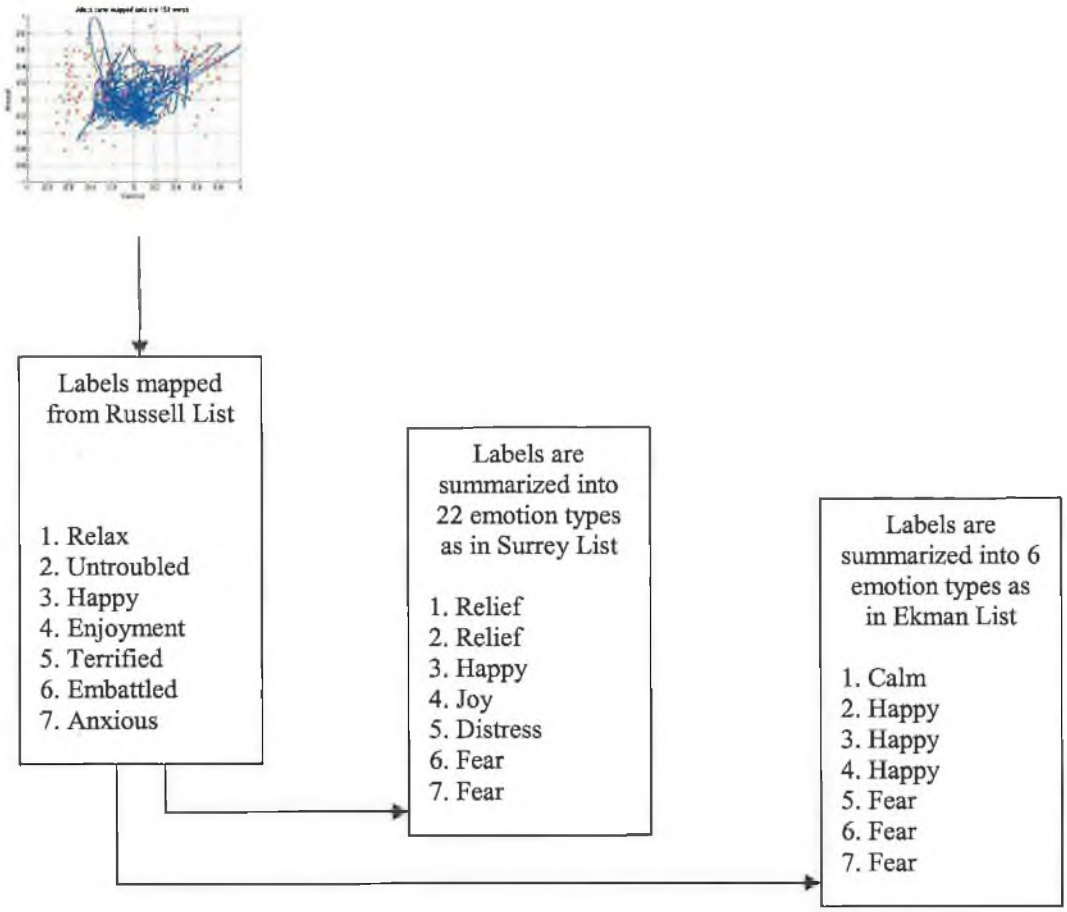


Figure 6-19. Three list of labels of varying granularity in terms description.

The labels are then summarized to form two additional lists of labels that are more general in description that are used to explore the granularity of affective labeling in the experimental evaluation in Section 7.6 (Results and Discussion of the User Evaluation). The second list is made up of 22 emotion types that were suggested for use by [Salway and Graham, 2003], where it is referred to as the "Surrey List" (based on the author's university name). The third list is even broader and expresses only 6 emotion types that are based on emotion recognition from facial gestures [Ekman and Friesen, 1978], referred to as the "Ekman List".

The Surrey List of 22 emotion types used by Salway and Graham [2003] was from the work of Ortony et al. [1988]. Salway and Graham [2003] expanded the 22 emotion types to a set of 627 emotion keywords using the WordNet ontology. For example, the emotion type "fear" has a number of keywords such as "alarmed". In order to make use of the Surrey List in this thesis, the 151 Russell keywords needed to be mapped onto the list of 22 emotion types. 50 of the 627 expanded Surrey keywords appeared in the Russell List, and so were easily mapped to the appropriate one of the 22 emotion types. The remaining 101 words were then manually mapped onto the emotion types using a standard English dictionary to identify the nearest emotion for each word. By associating each of the 151 words with one of the 22 emotion types, the affect curve will therefore output a list of labels expressing only 22 emotion types. This list of labels can then be plotted as shown in Figure 6-20 to show the dominant emotion types that are present in the video, similar to that found in Salway and Graham's [2003] work. Figure 6-20 shows the dominant emotions that are detected from a 10 minute segment of the movie "Road To Perdition". It can be observed that "fear" and "hate" are the dominant emotions as they appear the most often in the segment. The affect retrieval system automatically generates this list and a graph which can be a useful visual representation of dominant emotions.

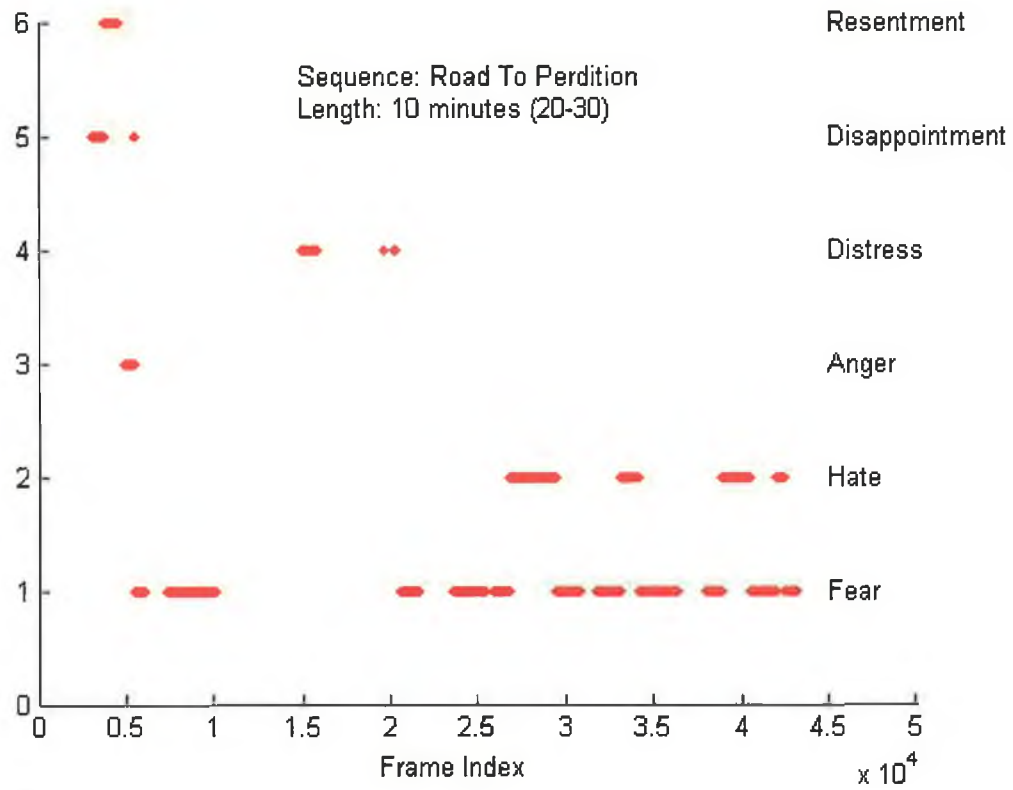


Figure 6-20. Plot of dominant emotion types detected from the Surrey List.

While the Surrey List was generated semantically, the Ekman List was generated using the positions of the VA emotion space. This is done by mapping the 151 original labels into 6 broad emotion types (happy, sad, surprise, anger, fear, and calm) by clustering their positions as shown in Figure 6-21. The 151 labels that fall within each boundary are associated with the corresponding emotion type. The clustering was done manually and based on intuition on where these 6 emotion types would fall on the VA emotion space while studying the distribution of the 151 original Russell's list of labels. Thus this list gives the broadest general description of the affective content contained in a video.

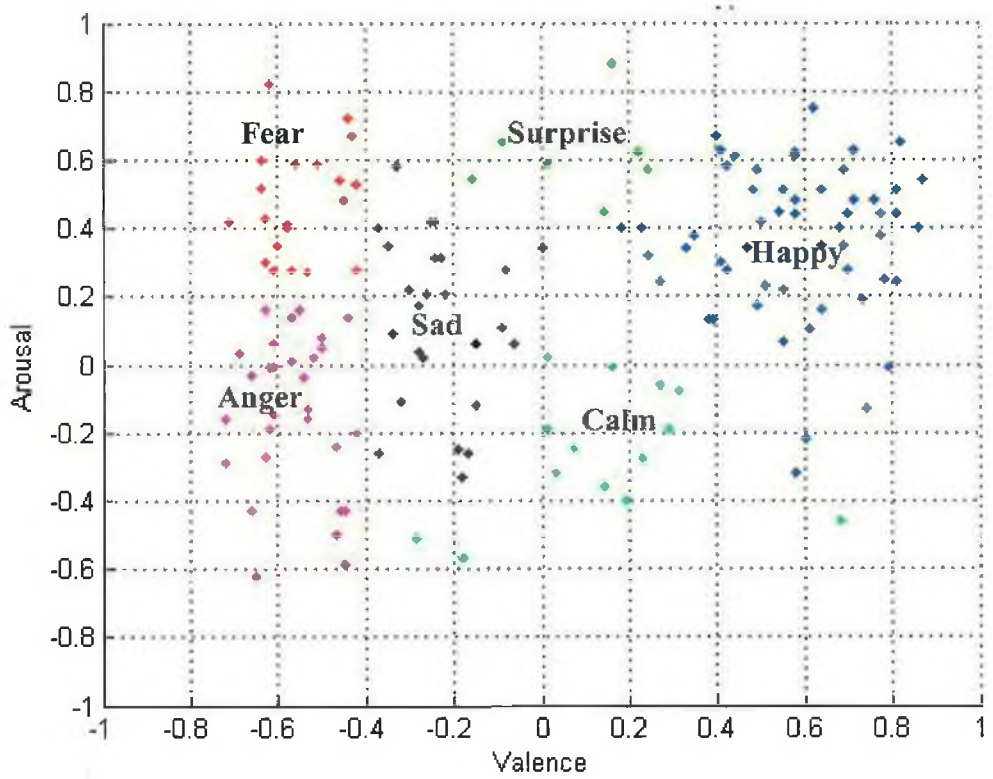


Figure 6-21. Distribution of Ekman's 6 emotion types.

6.8 Information Retrieval

The extracted verbal labels, derived from the original Russell List are then fed into the IR component of the system. This enables users of the system to search for documents with a desired emotional content by using ad hoc IR methods. The IR system uses the Okapi BM25 model as described in Section 5.4 (Combining the Evidence).

For each video that was processed, the detected verbal labels are split into 5 minute segments or "documents" (a text IR term). Each of the documents can then be treated as containing a number of emotional keywords, mirroring text documents that contain frequently occurring keywords that might indicate some significance.

Thus the search terms or verbal labels are weighted using the Okapi BM25 combined weight (CW). The Okapi BM25 tuning constant K1 was set to 1.5 and b was set to 0, as the documents are all the same length and no document length compensation was required. The matching score for each document is computed by summing the

weights of terms appearing in the query and the document, which are then returned to the user in order of decreasing matching score. Presented with this sorted list, the users are able to select and play back the 5 minute segments of the video through the system for review.

6.9 Open-Vocabulary Query Mapping

The open-vocabulary query mapping component is implemented using the WordNet::Similarity Perl module that traverses the WordNet ontology to give a measure of relatedness between two words using the Hirst and St-Onge's algorithm (Section 5.7.2 Measures of Relatedness). Given two words, it will output a score between 0 and 16, where 0 means it couldn't find a path between the two words, and 16 if the path is very short.

Therefore when a user's query is fed into the system, it will call up WordNet::Similarity to compare each of the query words to the original 151 labels from the Russell List. If a word is highly related, it will score higher. User query words that are found in the Russell List are awarded the maximum similarity score of 16.

Six different query processing strategies were explored in the known-item search task in Section 7.6.7 (Known-Item Search Task). The first four strategies uses the open-vocabulary query mapping component to map the user query words to the system's Russell list of 151 labels to form the final query that were fed into the IR component for retrieval. The first four strategies were "Unweighted full expansion", "Weighted full expansion", "Unweighted best expansion", and "Weighted best expansion". The remaining two are the "Reweighted" and "Bypass" strategies.

Full expansion means that all labels that have relatedness scores above 0 for each user-supplied query word were taken and fed into the IR component. This in effect fully expands the user query to the most number of labels, hence "full expansion".

Best expansion means that only the labels with the highest (best) relatedness score for each user query word were taken and used for the IR component. Sometimes a query word will have 2 labels with the highest relatedness score (such as 8 and 8), therefore

both labels are taken and fed as query into the IR component. Best expansion is more selective, therefore less labels were taken and fed into the IR component.

Unweighted or weighted determines whether the final query of labels in the IR component were weighted according to the relatedness scores that were derived earlier. For the unweighted strategy, every query label that has a relatedness score were said to be equally related, so the labels is fed into the IR component to generate a ranked list. For the weighted strategy, each query label's relatedness score were multiplied with the combined weights of the Okapi BM25 IR model in the IR component to generate a ranked list. This has the effect of giving labels with higher relatedness scores more emphasis in the IR process.

The subsections below shows the mathematical formula of deriving the combined weights CW(i,j) according to each strategy:

- **Unweighted full expansion**

For the "Unweighted full expansion" strategy, the combined weights of the Okapi BM25 IR model CW(i,j), where i is the query term, can be obtained,

$$CW(i,j) = \sum_{i_R} cw(i_R,j)$$

If the relatedness score $rel_{HS}(i_R,j) \neq 0$, where i_R is the relevant term.

- **Weighted full expansion**

For the "Weighted full expansion" strategy, the combined weights of the Okapi BM25 IR model CW(i,j), where i is the query term, can be obtained,

$$CW(i,j) = \sum_{i_R} rel_{HS}\,(i_R,j) \times cw(i_R,j)$$

If the relatedness score $rel_{HS}\,(i_R,j) \neq 0$, where i_R is the relevant term.

- **Unweighted best expansion**

For the "Unweighted best expansion" strategy, the combined weights of the Okapi BM25 IR model CW(i,j), where i is the query term, can be obtained,

$$CW(i,j) = \sum_{i_R} cw(i_R,j)$$

If the relatedness score $rel_{HS}\,(i_R,j) = max$, where i_R is the relevant term.

- **Weighted best expansion**

For the "Weighted best expansion" strategy, the combined weights of the Okapi BM25 IR model CW(i,j), where i is the query term, can be obtained,

$$CW(i,j) = \sum_{i_R} rel_{HS}(i_R,j) \times cw(i_R,j)$$

If the relatedness score $rel_{HS}\,(i_R,j) = max$, where i_R is the relevant term.

The fifth strategy called "Reweighted best expansion" is a modification of the "Weighted best expansion" strategy. The relatedness scores were first raised by an exponent value before it was multiplied with the combined weights of the Okapi BM25

IR model. This gives the higher range of the relatedness scores heavier weights and emphasis.

- **Reweighted best expansion**

For the "Reweighted best expansion" strategy, the combined weights of the Okapi BM25 IR model CW(i,j), where i is the query term, can be obtained,

$$CW(i,j) = \sum_{i_R} rel_{HS}(i_R,j)^X \times cw(i_R,j)$$

If the relatedness score rel_{HS} (i_R,j) = max, where i_R is the relevant term and X is the exponential value.

The sixth strategy called "Bypass" is to bypass the open-vocabulary query mapping component of the system and directly feed the queries into the IR component to generate a ranked list, therefore any query words not found in the system's Russell list of 151 labels is simply ignored.

- **Bypass**

For the "Bypass" strategy, the combined weights of the Okapi BM25 IR model CW(i,j), where i is the query term, can be obtained,

$$CW(i,j) = \sum_{i_R} cw(i_R,j)$$

If query term i is found in the list of the 151 Russell list of labels, where i_R is the relevant term.

6.10 Summary

This chapter has presented the design and explanation of the affect-based video indexing and retrieval system proposed in this thesis to investigate affect-based retrieval of video. The prototype system was developed using Matlab along with other downloaded software such as WordNet, WordNet::Similarity, VLC media player and block-matching algorithm scripts from Matlab's website.

The system was divided into 5 components to perform feature extraction, valence and arousal modeling, verbal labeling, open-vocabulary query mapping and information retrieval. For feature extraction, 6 low-level audio and visual features are extracted, smoothed and combined to model the affective features valence and arousal. Valence and arousal are then combined to form the "affect curve" which is used for verbal labeling using the Russell list of 151 pre-defined emotional verbal labels or words. The associated verbal labels are used to index the videos.

These labels are then summarized into 22 emotion types called the "Surrey List" and 6 emotion groups called the "Ekman List". When a user inputs a text query, the affect-based system maps the words that are not found onto the pre-defined Russell list of emotional labels using a measure of similarity between the query word and the keywords on the Russell List using WordNet::Similarity. This enables the user to use they wish regardless of whether they appear in the Russell list indexing vocabulary of the system. The query is then processed using the Okapi BM25 IR model to retrieve a ranked list of videos that are the highest scoring relative to the query.

Chapter 7 Experimental Investigation

7.1 Introduction

The previous chapters of this thesis presented the necessary literature review and the design considerations of an affect-based video retrieval system. This chapter presents an experimental investigation of this novel affect-based video indexing and retrieval system for the retrieval of films. The test data gathering activities and experiment procedure are presented, followed by experimental analyses and conclusions.

The experimental investigation presented here is designed to explore the behaviour and potential of affect-based video indexing and retrieval. This involves evaluating its effectiveness as a system for locating relevant content based on affective features. In order to do this, human volunteers who had never used or been involved in the development of this system were used in the evaluation to reduce bias. The three main questions addressed in this experimental investigation are:

- How accurate is the system at detecting affective features?
- Can the system retrieve results using users' description of emotional content?
- Does the system retrieve relevant results?

In order to answer those questions, the experiment was divided into three tasks that the volunteers were required to complete. The first task required the volunteers to rate the accuracy of the system by comparing the video and its detected affective features. The second task required them to generate textual descriptions of emotional content in a piece of video, which were then used as a query to test how well the system could locate relevant pieces of video in emotional content from a large database based on this query. The third task required the volunteers to review a list of retrieved results from the system and rate whether the results could be considered relevant to their query. After the volunteers had completed the experiments and left, a known-item search task was performed to retrieve clips from the database using the clips' queries. This task was done

to explore how effective the system is at retrieving the film segment used to generate the request and to what extent it can be fine-tuned by using different setups and parameter values to achieve the best retrieval results.

7.2 Experimental Test Data

It is expected that the most immediate and obvious application of an affect-based video retrieval system would be for browsing through a large database of films. Hollywood-style films represent a compelling source of video data especially for an affect-based system due to the richness of emotional content within them. In addition, emotional content is much more pronounced in films than other videos such as TV news, therefore making it easier to derive what emotions the film is trying to project. This has the added advantage of making it easier for human viewers to rate how well the system is performing in the evaluation. Therefore this experimental investigation of the system takes a set of these films as its video data. The films were gathered from the CDVPlex project [Rothwell et al, 2006] that involved human subjects watching a variety of films and having their physiological responses monitored and recorded as well as providing written feedback on the highlights of the film and film watching preferences.

The films that were gathered from the CDVPlex project come in 2 separate files for each film, one with only the audio stream, and one normal movie file that contains both video and audio. The visual stream comes from the normal movie file with an .avi extension, using the compressed DivX video codec [DivX] which is based on the MPEG format. The audio stream comes in a file with a .wav extension, an uncompressed audio data file. For movie playback or review, the .avi file with both audio and video was used. Each file was named after the title of the film, therefore the film "Nora" will have the filename "Nora.avi" and "Nora.wav".

In order to test the system's effectiveness in retrieving relevant results, a suitably large corpus of video data had to be used. A total of 39 films from the CDVPlex project were processed covering a wide range of genres from action films to comedy and horror films. This amounts to approximately 80 hours of video data, which is comparable to a large-scale video retrieval system evaluation such as [TRECVID]. Each film was then

processed to extract visual and audio features using a bank of 10 desktop computers running on Intel Pentium 4 2.8 GHz processors with 512 MB of random access memory (RAM). Processing took a total of two weeks, but the long period of computation is due to unoptimized code running in the Matlab development environment. While not a priority in this thesis, the length of time needed to process these features can be significantly reduced through code optimization and compiling the code into binary executable files. In order to facilitate video browsing and retrieval, each film was broken up into 5 minute segments, giving a total of 939 segments. The segments are all equal lengths of 5 minute segments that were split without any overlaps. The 5 minute length was chosen based on the reasoning that it would be sufficiently long enough to allow users to be able to observe and rate the dominant emotional content of the clip without requiring an excessive time commitment to complete the interactive experimental sections.

7.3 Experiment Procedure

The experiment was divided into two sessions to avoid user burnout as well as provide flexibility for scheduling. This is because the first session takes about one hour to complete, and the second session takes two hours to complete for a total of three hours. The users worked at a prepared computer, and were guided through the process every step of the way by the experimenter. They were required to complete an informed consent form to state that they had understood the experiment's purpose and methods, and what was expected from them. They were also briefed on some general detail of the valence and arousal curves, as it was vital they understood the difference between the two before any the experiments started. A very important point to maintain data consistency was to explain to every user clearly that the system is detecting affective features (emotion that the scene is trying to project to the viewer) as opposed to what the viewer emotions actually are or the film character's emotions. They were also informed that they were allowed to withdraw from the experiment at any time or take breaks. A blank copy of the complete set of forms completed by each volunteer during their participation in the experiment is given in Appendix B (Questionnaire).

7.3.1 First Session

The first session of the experiment began with the user filling in their particulars in a feedback form. They were then asked to choose five films most familiar to them from among the 39 films that have been processed. They then rated how well they remember the films in the feedback form.

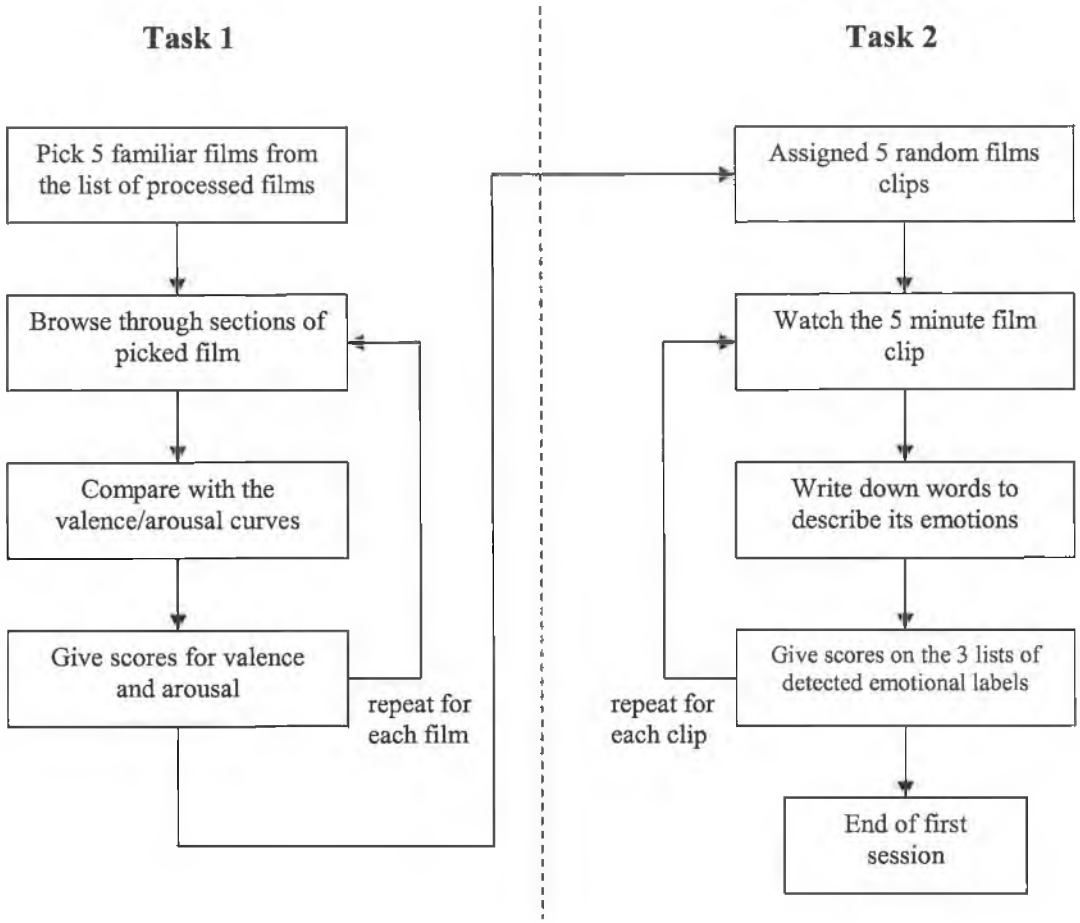


Figure 7-1. First session of user evaluation experiment.

Figure 7-1 shows a summary of the first two tasks. The first task is to rate the quality of the valence and arousal curves that were automatically extracted from the films. A movie player starts playing the first of the five films that have been chosen by the user. Below the movie player in the bottom half of the monitor a plot of the valence

curve is displayed. The timeline on the movie player is lined up with the valence curve, so the user can drag the movie player timeline to any part of the movie according to what they see on the valence curve. The user is then able to quickly browse through the movie to see if the happenings in the film correspond to the valence curve. In general they were told to see if the trend of the valence curve matches the film scenes. The users were told not to spend more than 5 minutes on this task, although they were free to exceed this if they needed to scrutinize the valence curve further. When they were done, they wrote down on the feedback form a score from 0 to 10 to indicate if they agree the valence curve is correct in detecting the valence components of emotions conveyed by the curve, as interpreted by the user. When they were done, they closed the movie player and a timer recorded how long they took to browse and give a score. At this moment the movie player starts again with the same movie, but with an arousal curve and the user was asked to score the arousal score in a similar way. This process is then repeated for each of the five movies they had selected.

The second task was to watch 5 film clips and choose words to try to describe dominant emotions contained in the clip for a query and retrieval task. A random set of five of the 5 minutes film clips was chosen from the 39 processed movies, and assigned to the user.

They were shown the first clip and allowed to repeat or browse through the clip as many times as they wanted. Once they finished watching the clip, they were asked to write down on the feedback form words that describe the dominant emotions contained in the video clip. They were allowed to use as many or as few words as they liked, and use any words they wanted. The only restriction was that the words should describe the emotional content and not the plot or character information in the film clip.

After this, 3 short lists of the emotional verbal labels that had been automatically detected using the system were shown to the users. The three short lists consist of the top few labels that occurred the most number of times in the clip. The first of the short list had the top 10 labels occurring from the Russell list, the second list had the top 5 labels occurring from the Surrey list, and the third list had the top 3 occurring labels from the Ekman list of labels. Therefore each list represents a description of the emotional content of the film clip in varying granularity (Russell's list of 151 verbal labels is the most

detailed, followed by Surrey's list of 22 verbal labels and Ekman's list of 6 verbal labels). Users were then asked to give a score from 0 to 10 of how much they agreed that the list of detected emotional verbal labels described the film clips.

This process was then repeated for each of the 5 film clips where a timer recorded how long the user needed for each clip. Once they were done, the first session was finished and they were thanked for their time and allowed to leave.

7.3.2 Second Session

For each clip, the words that were written down by the user were treated as a text query and fed into the retrieval system by the experimenter to retrieve the top 5 scoring (result) film clips that is detected by the system to be relevant to the query. For each result of the film clips, the 3 short lists of emotional verbal labels from the Russell, Surrey, and Ekman lists were automatically detected using the system as well.

The second session involved asking the user to rate the results of the system in retrieving videos with similar emotional content to the 5 (original) film clips that were assigned to them in the first session. For the ratings, users had to rate how similar the result film clips were to the original film clip in terms of emotional content on a scale from 0 to 10. In addition, the users were asked to give a score from 0 to 10 for each of the 3 short lists of the result film clip's detected emotional labels. In short, the users were always comparing the retrieved result film clips and their detected emotional labels against the original film clips.

Task 3

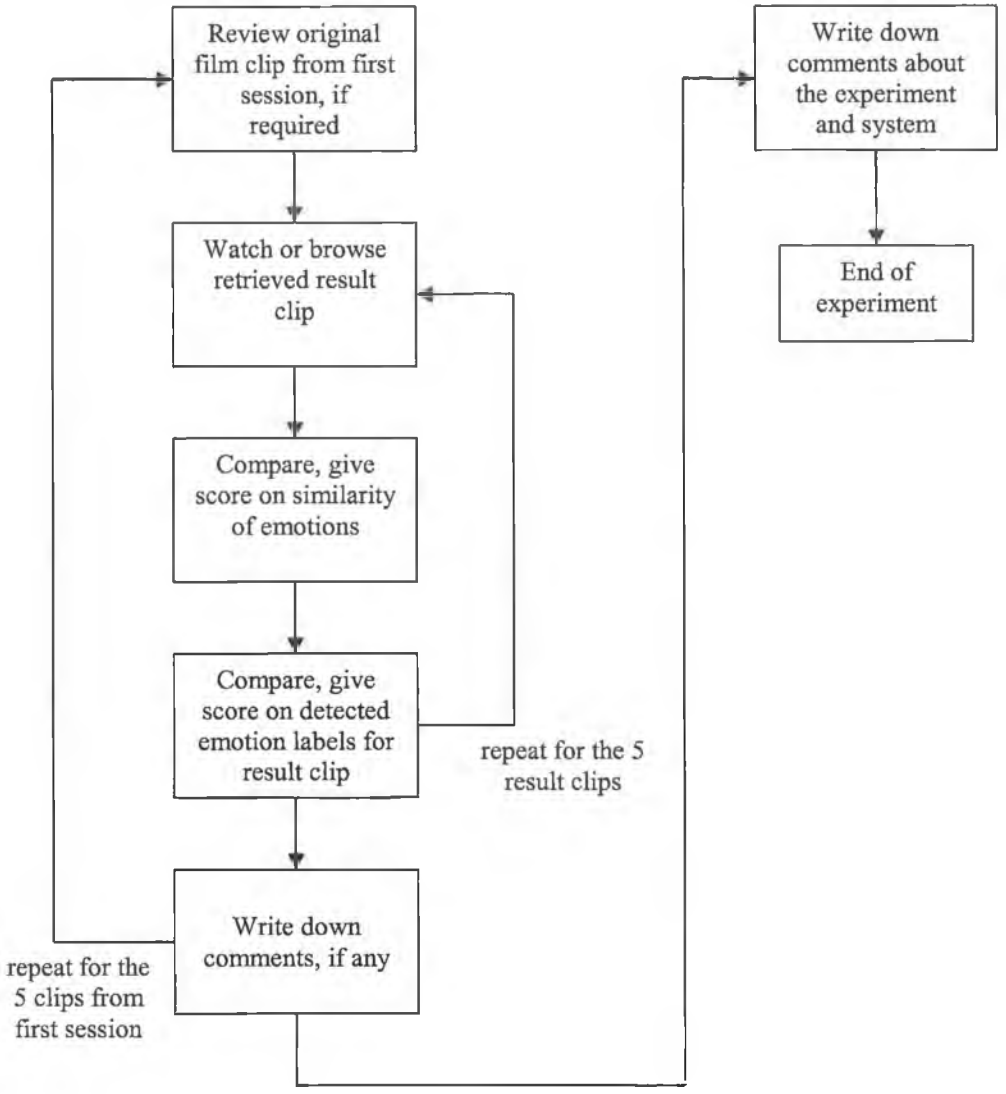


Figure 7-2. Second session of user evaluation experiment.

Figure 7-2 shows a summary of the third task in the second session. The experiment begins with the experimenter asking the users if they need to review the first of the five film clips that were randomly assigned to them in the first session. Some time is spent reviewing the clip if the user required it. Then the user is shown the 5 retrieval clips for the randomly assigned clip, one at a time. Users were allowed to watch through the entire 5 minute result clip or quickly browse through if they were familiar with it. After watching each clip, the user was required to think and write down their scores

before moving on to the next result clip. The experimenter recorded the time taken for the user to finish watching and rating each result clip.

This process was repeated for all 5 of the film clips that were randomly assigned to them, therefore giving a total of 25 result clips the user had to rate. Once this was done, users were asked to write down comments regarding the experiment and system, thanked again, and allowed to leave.

7.4 Human Volunteers as Users

A total of 8 volunteers agreed to participate in the evaluation, which was split into two sessions to avoid burnout. The first session lasted for approximately one hour and the second session for approximately two hours. Where possible, the second session was conducted on the next day. The evaluation was conducted over a period of two weeks between the 8 volunteers, which were scheduled according to their availability. The 8 volunteers all speak the English language, are computer-literate and hold higher education qualification in computer science. Only one out of the 8 volunteers was female, and they had an average age range of 29. The information which the volunteers supplied were name, age, gender, how often they went to the cinema, what genre of films they liked, and how familiar they were with the films. From here on the volunteers will be referred to as users, and each of them assigned a number from 1 to 8.

Table 7-1. Users' cinema-going frequency.

Cinema-going frequency	Number of users
Seldom	1
1-2 times a year	3
Once a month	3
2-3 times a month	1
Once a week	0
More than that	0

As can be seen in Table 7-1, the users frequent the cinema on average between once a month and 1-2 times a year. However this does not include watching films on television at home or other media. This reveals that most of the users who participated do enjoy the film experience of the cinema.

In addition, the users' film genre preference was gathered. They were required to give a score between 5 (most preferred) to 1 (less preferred) for each genre of film.

Table 7-2. User's film genre preference.

User No.	Action	Drama	Horror/Thriller	Romantic	Comedy
1	3	2	4	1	5
2	3	5	1	4	2
3	2	5	3	4	1
4	4	3	2	1	5
5	3	2	4	5	1
6	2	5	4	1	3
7	5	3	4	2	1
8	4	2	5	3	1
Average total score	**3.25**	**3.38**	**3.38**	**3**	**2.38**

As shown in Table 7-2, by taking the average of the scores, the film genre preference of the group can be discovered. The total scores reveal that drama and horror/thriller films were the most preferred, followed by action, romantic and comedy. This preference could suggest that the users enjoy going to the cinema to experience the higher-quality visual and audio that a cinema provides to fully appreciate the horror/thriller, drama and action film genres.

7.5 Retrieval Evaluation Measures

In order to measure the system's performance, standard information retrieval (IR) evaluation measures are used. These evaluation measures compare system performance and provide evidence of any performance gain or loss arising from variations in retrieval techniques. The thesis is concerned with retrieval of film clips that exhibit similar dominant emotions, therefore a relevant result is a film clip that is assessed by a user to have similar emotional content to what was described in the query. As each clip is associated with a list of detected emotional labels similar to a text document, each film clip here is called a document (chapter 5 Information Retrieval and Document Matching).

- **Precision**

Precision is used to measure the percentage of returned documents that are relevant. Precision can be calculated by dividing the number of returned relevant documents by the total number of returned documents [Salton, 1983] [Van Rijsbergen, 1979]. The precision formula is,

$$\text{Precision} = \frac{\text{Number of returned relevant documents}}{\text{Total number of returned documents}}$$

If all the returned results are relevant, then the system is said to have achieved 100% precision.

- **Recall**

Recall is used to measure the percentage of available relevant documents that have been retrieved and is calculated by dividing the number of returned relevant documents by the total number of relevant documents [Salton, 1983] [Van Rijsbergen, 1979]. The recall formula is,

$$\text{Recall} = \frac{\text{Number of returned relevant documents}}{\text{Total number of relevant documents}}$$

When all relevant documents in a database have been retrieved by the system, then the system is said to have 100% recall. However, due to the difficulty of annotating films and the large amount of human effort required, which is also subjected to individual interpretation, recall values are not presented in this experimental investigation.

- **Mean-reciprocal-rank**

In a known-item search, the performance of a system can be obtained by looking at the rank at which the target document is retrieved. The mean-reciprocal-rank (MRR) is the mean of the reciprocal of the rank at which the known item was found, averaged over all the queries, and using 0 as the reciprocal for queries that did not retrieve the known document [Kantor and Voorhees, 1999]. Therefore,

$$\text{Mean-reciprocal-rank (MRR)} = \frac{1}{N} \times \sum_{i=1}^{N} \frac{1}{\text{answer}_i \text{ rank}}$$

where N is the total number of queries. The mean reciprocal is bounded between 0 and 1 (1 represents perfect retrieval), so its value can be easily understood without knowing how many documents were ranked. The mean reciprocal is equivalent to precision at 100% recall, as there is only one target document.

7.5.4 Difference Between Subjective and Standard Video Retrieval System Evaluations

A difficult issue of testing an affect-based video indexing and retrieval system is the issue of subjectivity and relevance. Standard evaluations of video retrieval systems are based on detecting clearly defined semantic units. Established work such as that found at [TRECVID] shows how evaluations are done by testing the feature detectors and user experiences with the system. Some examples of these are detecting faces, characters, events and objects. The result of the feature detectors therefore, can be distinctly separated into "yes" or "no" categories as long as the detection meets the criteria. For example, a face detector could therefore return a "yes" result if a face was detected, or "no" otherwise.

Emotional content however, is not so clearly defined. There are no immediately clear and distinct boundaries separating one emotion from another. And within each emotion, there can be varying levels of intensity in emotions. Furthermore, it is unknown

how much variance between people describing emotions can differ. One person's labeling of a scene as "happy" might differ from another that might choose to use a less strong description, such as "elated" or "hopeful".

Therefore a pragmatic approach is taken where volunteers are asked to judge whether the affective relevance of returned results is deemed acceptable. In this system's evaluation of relevance, users were asked to rate the system's performance on a scale ranging from 0 to 10, where the scores 0 to 3 means "disagree", 4-6 means "slightly agree" and 7-10 means "strongly agree". This approach eliminates any problems of bias that might be encountered through self-reporting of retrieval results by the author of this thesis. In addition, this approach which relies on subjective evaluations of the human users focuses the results on the usability of the system as determined by the users.

7.6 Results and Discussion of the User Evaluation

The following section gives a summary of the results of the 3 tasks, along with the known-item search task, analyses and any conclusions that can be made.

7.6.1 Task 1

For the first task, each user was asked to choose 5 films with which they are most familiar from the available list of 39 films, in order to rate the performance of the valence and arousal curves that were generated for the films. With 8 users, the total number of films evaluated was 40, but there were repetitions of films being evaluated, therefore only 20 of the 39 films were evaluated. This represents slightly more than half of the entire range of films that have been processed for use in the affect-based video indexing and retrieval system presented in this thesis. Two movies received the highest amount of evaluations, "Shrek" and "Titanic", 5 in each case, while other films were evaluated either three, two or one time, as shown in Table 7-3.

The decision to allow users to choose the films that they are most familiar with did not restrict the evaluation to only a small number of films. While the films "Shrek" and "Titanic" received more evaluations than others, the range of films that was

evaluated was still reasonable at 20 out of 39. This can be explained by the films' popularity, therefore more users have seen these movies before. Cult favourites such as "Leon", "Pulp Fiction", and "The Godfather" also make it among the top-picked films. Lesser-known films such as "Nora", "Roger and Me", and "Jackie Brown" were not picked. In addition, the age of the movie is also a factor, as older films such as the older "Star Wars" trilogy of films that were included in the list were not picked despite their immense popularity. This is because the users had not seen the films for a long time. Any future work that carries out this sort of experiment will have to take these factors into consideration so the users choose the best range of films from the available database for that experiment.

Table 7-3. List of films selected for arousal/valence evaluation.

Film Title	**Number of times selected for evaluation**
Shrek	5
Titanic	5
Finding Nemo	3
Pulp Fiction	3
The Godfather Part 1	3
Alien	2
Chariots of fire	2
ET	2
Harry Potter and the Philosophers Stone	2
Leon	2
Minority Report	2
Chocolat	1
Disco Pigs	1
Harry Potter and the Chamber of Secrets	1
Indiana Jones and the Last Crusade	1
Indiana Jones and Raiders of the Lost Ark	1
Pirates of the Caribbean	1
The English Patient	1
Lord of the Rings: Fellowship of the Ring	1
Toy Story	1

In addition to asking users to choose the five films with which they were most familiar, they were also required to score how well they remember each of the films they had chosen. The three categories that they could choose were: Remember most scenes, Remember some scenes, Remember plot film only, or Not familiar with the film. This is

to show the degree of familiarity that the users exhibit with the films that they have chosen. This is important as the task of rating the accuracy of the valence and arousal curves' requires some understanding and memory of the film's general structure and plot. This further justifies the experiment design which was selecting films for evaluation rather than allocating films in an even distribution, since the users must already be at least somewhat familiar with the films in order to review the quality of the valence and arousal curves.

Table 7-4. Users familiarity with their chosen five films.

User No.	Remember most scenes	Remember some scenes	Remember plot of film only	Not familiar with the film
1	4	1	0	0
2	5	0	0	0
3	2	3	0	0
4	2	3	0	0
5	2	2	1	0
6	3	2	0	0
7	2	1	2	0
8	2	2	1	0
Total score	**22**	**14**	**4**	**0**

Table 7-4 shows the user's familiarity with their selected films. The results show that all the users had seen and at least remembered the plot of all the films they had chosen. On average the users had a vivid understanding and memory of the film they had chosen, based on the number of times they chose the "Remember most scenes" and "Remember some scenes" category, therefore hopefully giving the results of their assessment of the valence and arousal curves a high degree of reliability.

After this information had been collected about the chosen films, the users were required to rate each film according to how accurate the valence and arousal curves were for the entire film. The users gave a score for the two curves for each film, giving a total of 10 scores that ranged from 0 to 10 (0-3 = disagree, 4-6 = slightly agree, 7-10 = strongly agree). The scores are tabulated in Table 7-5 and 7-6, along with their averages, with film 1, 2, 3, 4, and 5 not necessarily the same films for each user.

Table 7-5. Arousal accuracy for each user and the five films they've chosen.

Arousal accuracy for user:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Film 1	7	8	6	8	9	7	10	8
Film 2	9	7	7	7	8	9	10	8
Film 3	6	6	6	6	7	5	9	9
Film 4	8	7	5	7	8	8	9	8
Film 5	9	5	7	6	9	9	10	9
Average scores	**7.8**	**6.6**	**6.2**	**6.8**	**8.2**	**7.6**	**9.6**	**8.4**

Table 7-6. Valence accuracy for each user and the five films they've chosen.

Valence accuracy for user:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Film 1	5	6	4	5	8	4	9	6
Film 2	7	6	4	3	6	5	9	7
Film 3	8	7	3	3	7	7	9	6
Film 4	6	6	5	4	6	5	8	8
Film 5	8	6	3	3	8	6	10	8
Average scores	**6.8**	**6.2**	**3.8**	**3.6**	**7**	**5.4**	**9**	**7**

Tables 7-5 and 7-6 show the valence and arousal accuracy scores that the user assigned for each of the user-picked films. The lowest score for the arousal curves is 5 and the highest is 10. For valence, the lowest score was 3 and the highest was 10. The scores reveal that arousal outperformed valence curves in most instances.

In addition, it can be observed from these tables that for each user, the difference in the average scores between valence and arousal was small. This small difference can be observed across all users except for users 3 and 4, where the difference was the biggest. Overall, based on the opinions of the users, arousal appears to be captured more accurately than valence. This is consistent with the existing view that capture of valence is more difficult than arousal [Kok, 2006]. However, the general similarity of the results here indicates that this system is extracting valence with reasonable accuracy.

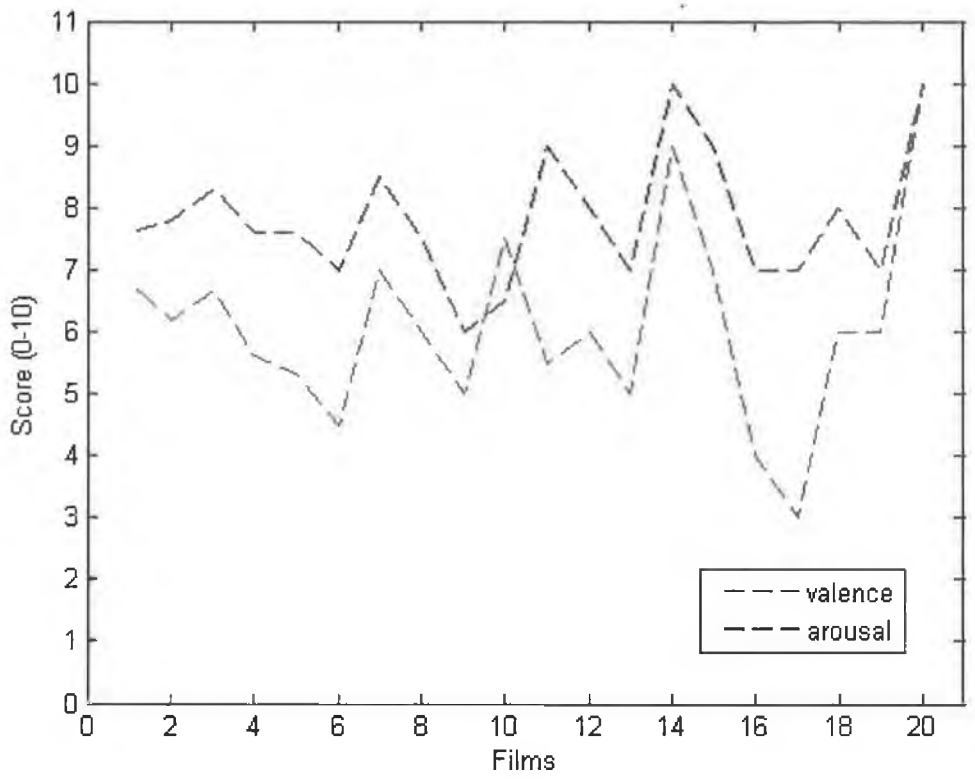


Figure 7-3. Accuracy scores of valence and arousal for user-picked films.

Figure 7-3 shows the valence and arousal accuracy scores for the 20 films that were picked out of the 40 available films. The average of the scores was taken for films that were picked more than once. As can be observed, the average arousal scores were quite high and concentrated in the upper regions of the scale. The valence average scores are more evenly distributed across the middle and upper region of the scale. Although the average valence scores are lower for all films, compared than those for arousal, it does not fall too far behind except for film 17 "Pirates of the Caribbean" where the variance was the highest. In this instance, the score for valence were quite low at 3 when compared to the other scores. Only in the case of film 10 "Leon" the average valence score was higher than the arousal score. However, each of the movies "Titanic", "Harry Potter and the Philosophers Stone", and "Pulp Fiction" also had one instance where the valence score were higher than the arousal score. Therefore 4 out of the 40 valence scores given by users were higher than the arousal scores.

Table 7-7. Comparison of valence and arousal (V/A) scores for films that were picked three or more times.

User	Shrek (V/A)	Titanic (V/A)	Finding Nemo (V/A)	Pulp Fiction (V/A)	The Godfather Part 1 (V/A)
1	8/9			6/8	
2		6/5			
3	5/5	3/7	4/7		
4				3/6	3/6
5	4/7				8/9
6		6/9		7/5	5/8
7	9/9	8/9	9/10		
8	8/8	8/9	8/9		

Table 7-7 shows the arousal and valence scores for 5 films that were picked by users three or more times. The films had some consistent scores between the different users. This gives some indication of the consistency of the scores between different users. The valence and arousal scores can be grouped into high, medium and low scoring pairs. It can be observed that the film "Shrek" had three pairs of high valence and arousal scores that were either 8 or 9. The remaining two pairs had medium scores that ranged from 4 to 7. The film "Titanic" had two pairs of high scores that were either 8 or 9. This is followed by the remaining three pairs of medium scores that ranged from 3 to 9. The film "Finding Nemo" had two pairs of high scores that ranged from 8 to 10 and a lower pair of scores that were either 4 or 7. The film "Pulp Fiction" had two pair of medium scores that ranged from 5 to 8, and a lower pair of scores that were either 3 or 6. The film "The Godfather Part 1" had one pair of high scores that were either 8 or 9, one pair of medium scores that were either 5 or 8, and one pair of low scores that were either 3 or 6.

Table 7-8. Global average scores (precision) for valence and arousal.

	Global average score (Precision)
Arousal	**7.7**
Valence	**6.1**

The precision of the system in detecting the valence and arousal curves can be obtained by taking the global average score that had been given by the eight users. When

the scores are averaged over all movies, the global average score for arousal and valence is 7.7 and 6.1 respectively, as shown in Table 7-8.

7.6.2 Discussion on the Results of Task 1

On average the arousal scores were higher than the valence scores. This result is expected as arousal is better understood and current research such as in sports video summarization uses the same low-level features to achieve high precision values in selecting good highlight points for the summary. The average score of 7.7 presents a very strong case for further research work in detecting affective features. It shows that affective features can be detected and their performance has been deemed to be working well by multiple human users.

While arousal is working well, the average score of 6.1 for valence makes an even stronger case that emotional content can be detected. This shows that multiple human users agree that affective features detected from the films do work and are able to show the trends, development and progression of feelings. The relatively small difference between arousal and valence scores indicate that the valence curves performed quite well compared to the more established arousal. This is in spite of the fact that the valence modeling presented in this system deviates from Hanjalic and Xu's [2003] model by calculating the pitch values for the entire length of the audio without any separation into voiced or unvoiced segments. In addition, the results here also justify the decision to incorporate Kok's [2006] preliminary findings that colour information could be used to model valence into the system. Valence is a relatively new feature and not as well understood as arousal; therefore it is likely that this performance could still be increased further through the selection of more discriminating low-level features.

The results presented are also the first attempt at quantifying the performance of the valence and arousal modeling. Previous justifications for valence and arousal modeling relied on presenting the plot of the curves and describing the scenes in detail, which could be subjected to bias or subjective interpretation [Hanjalic and Xu, 2003] [Kok, 2006]. Due to such an evaluation method, it is difficult to tell if any improvement such as the selection of low-level features constitute an improvement in the valence and

arousal detection. This was one of the major problems facing the development of the affect-based indexing and retrieval system in this thesis as there were no baseline results and experimental guidelines to follow that could indicate an improvement over previous similar work, e.g. that of Hanjalic and Xu [2003] and Kok [2006].

Based on multiple human evaluations, the results provide strong evidence of the system's ability to detect affective features. The approach here to use different human users with minimal knowledge of the VA emotion space and a 0 to 10 scoring scale reduces the bias and subjectivity and contributes a baseline figure and methodology for future work.

7.6.3 Task 2

For the second task, users were required to write down textual descriptions to describe the emotional content for each of the 5 film clips that were randomly assigned to them, as well as giving a score for how well the verbal labels detected by the system describe the emotions contained in the film clips. In later experiments, the user's textual description of the films was used as an IR query to the system to attempt to retrieve the clip itself from the database of film clips, but also to find clips with similar emotional labels.

Table 7-9. Number of words each user used to describe emotions for each clip.

User	Clip 1	Clip 2	Clip 3	Clip 4	Clip 5	Total number of words	Average number of words per clip
1	5	3	3	4	3	18	**3.6**
2	9	8	14	8	9	48	**9.6**
3	3	2	0	3	3	11	**2.2**
4	3	4	6	4	4	21	**4.2**
5	3	5	5	7	8	28	**5.6**
6	6	8	7	6	5	32	**6.4**
7	3	3	4	7	3	20	**4**
8	6	8	9	9	4	36	**7.2**

It was observed that different users entered a widely varying number of words to describe the emotional content for each of the film clips. As shown in Table 7-9, user 2

used an average of 9.6 words, which is the highest number of words, while user 3 used an average of only 2.2 words, which is the least number of words. There was only one case where the user (user 3, clip 3) was unable to provide any words to describe the emotions in the clip. The feedback from user 3 revealed that the film clip in question did not express any dominant emotion and is made up of a montage of slow moving and neutral scenes. Thus user 3 could not think of any words to describe the clip from an affective perspective and left it blank. In this particular case, user 3 was allowed to use the top 10 labels that were detected by the system for this clip as it's textual description instead. On average the users entered 5 words to describe each of the film clips.

From among the words that the users wrote down to describe the emotional content of the film clips, there were a lot of words that were not contained in the 151 words or labels in the Russell list with defined arousal and valence values. While these labels describe a wide range of emotions, it was observed that users typically used a far wider range of words to describe emotions. Table 7-10 shows the percentage of words that were not found in the Russell list, ranging from 75.0% to 36.4% for different words. Among some of the words that were not found in the system were; "trepidation", "nostalgic", "mundane", and "mysterious".

Table 7-10. Number of affect description words for film clips entered by users that are found and not found in the system.

User	Total words used by the users	Words found in system	Words not found in system	Percentage of words not found in system (%)
1	18	5	13	**72.2**
2	48	12	36	**75.0**
3	11	7	4	**36.4**
4	21	6	15	**71.4**
5	28	8	20	**71.4**
6	32	10	22	**68.8**
7	20	8	12	**60.0**
8	36	15	21	**58.3**

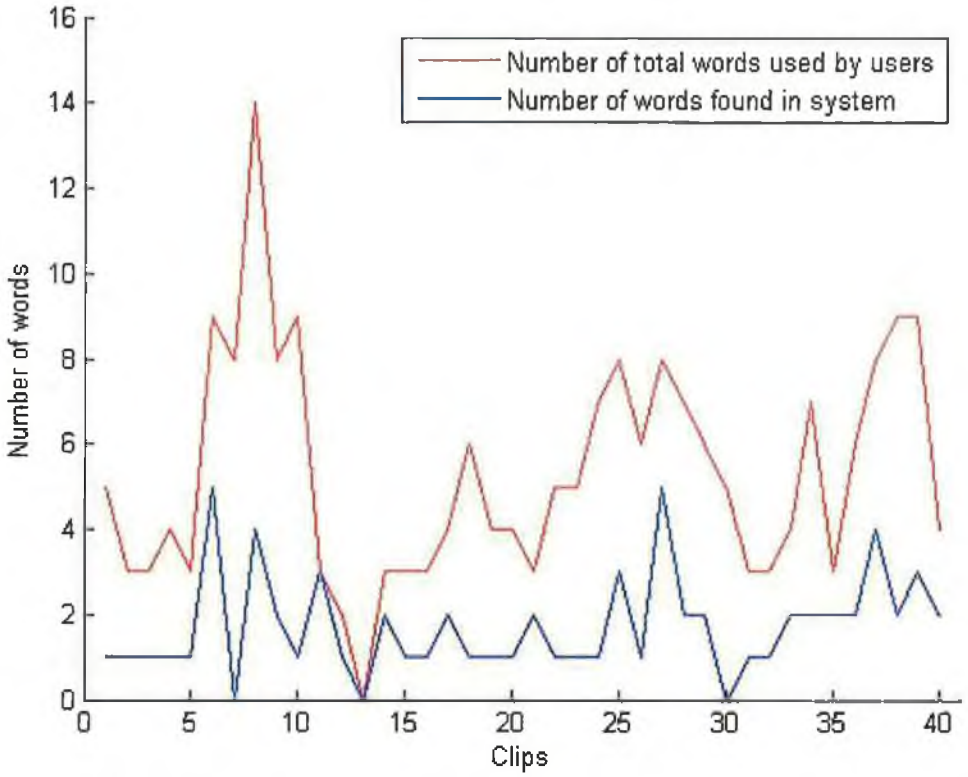


Figure 7-4. Total number of words users entered that are found and not found in the system.

Figure 7-4 shows the gap between the total number of words entered by the users and the words contained in the system vocabulary for each of the 40 film clips described by the 8 users. Except for two instances where all the words used were found (clip 11 and 13), all the clips had a significant gap. This shows that users, in order to describe the emotional content of a film clip had consistently used words that were not found in the system.

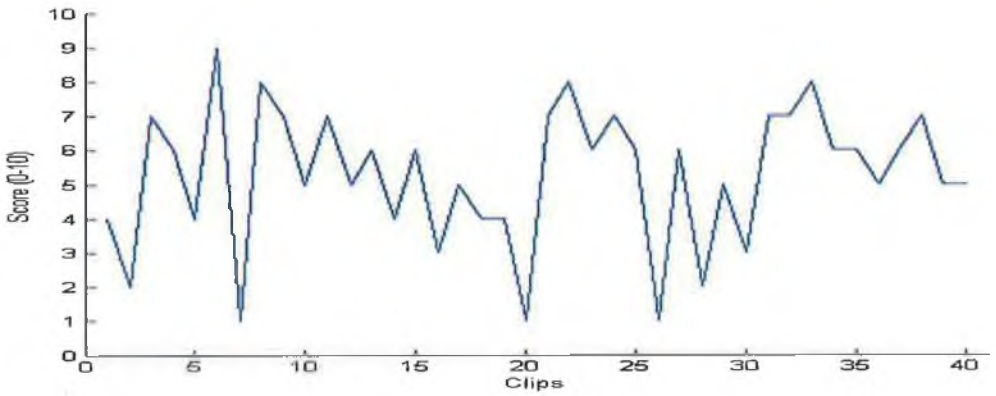


Figure 7-5. Scores for each clip's detected emotional labels from Russell list.

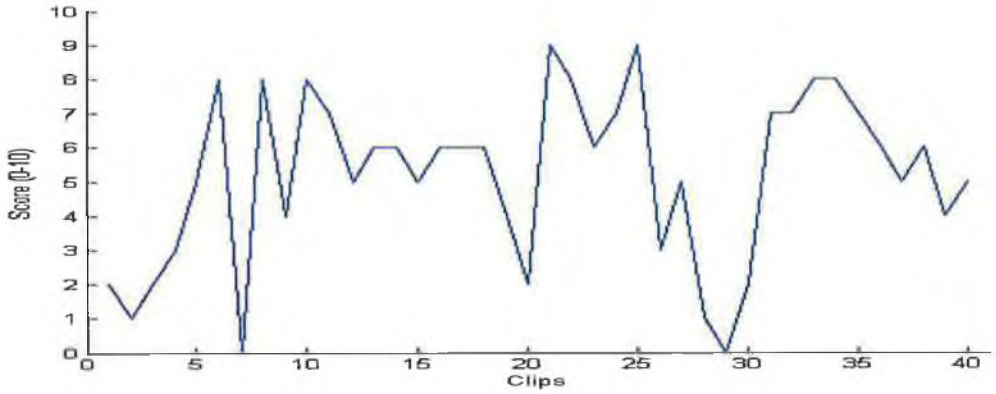


Figure 7-6. Scores for each clip's detected emotional labels from Surrey list.

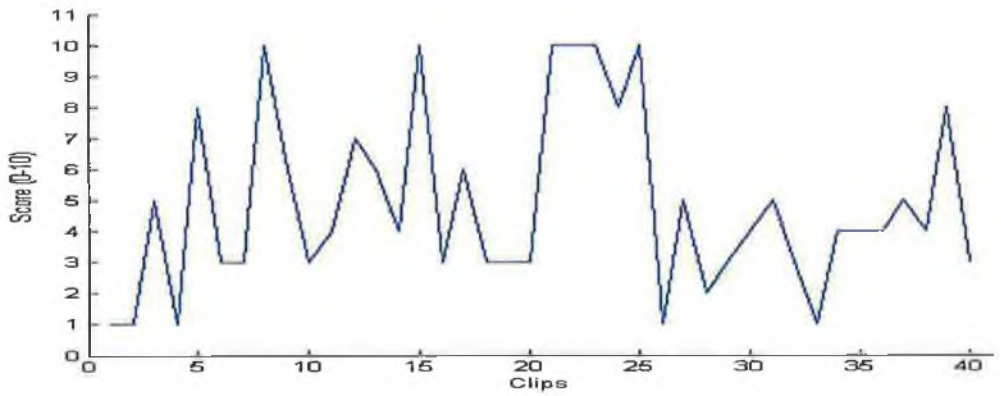


Figure 7-7. Scores for each clip's detected emotional labels from Ekman list.

Figures 7-5, 7-6 and 7-7 show the results of the user's assessments of the 3 lists of emotional verbal labels that had been automatically detected by the system for each of the assigned clip. For each list users were asked to assign a score from 0 to 10 to indicate how well they felt the list of labels describe the affective features of each assigned clip. This was repeated for all clips. It can be observed from the figures of all 3 lists that the scores are spread across the whole scale of 0 to 10. User 5 in particular (clips 21 to 25) assigned the highest scores across the three lists.

In Figure 7-5, the Russell list scores exhibited scores that were more evenly distributed across the scale from 2 to 9. In the Surrey list scores as shown in Figure 7-6, except for the two instances where the scores were zero (clip 7 and 29), the line is concentrated around the 4 to 8 range of the scale. However, in contrast to this, the Ekman list scores as shown in Figure 7-7 exhibits a more "peaky" line, with the scores concentrated around the 1 to 5 range of the scale.

Table 7-11. User-assigned average scores of detected emotional labels for each list.

Average scores per user	**Russell list**	**Surrey list**	**Ekman list**
User 1	4.6	2.6	3.2
User 2	6	5.6	5
User 3	5.6	5.8	6.2
User 4	3.4	4.8	3.6
User 5	6.8	7.8	9.6
User 6	3.4	2.2	3
User 7	6.8	7.4	3.4
User 8	5.6	5.2	4.8
Global average scores for each list	**5.28**	**5.18**	**4.85**

This observation can also be seen when the values are averaged as shown in Table 7-11. Looking at the averaged scores for each list, it is found that the Russell list's score of 5.28 slightly outperforms the Surrey list's score of 5.18, which is then trailed by the Ekman list's score of 4.85. In addition, looking at the average scores given by each user, except for user 3 and 5, the scores are lower as the labels get more general (Russell has 151 labels for the system to choose from, Surrey has 22 labels, and Ekman has only 6

labels). Therefore the system can be said to be achieving at overall labeling accuracy slightly above 5.

7.6.4 Discussion on the Results of Task 2

The objective of asking users to write down words to describe the emotional content of the film clips was to explore what words users would typically use in an IR task. While the system only has a vocabulary of 151 words or labels from the Russell list, it was expected that users would use a far wider range of words.

Table 7-9 revealed that users typically use around 5 words to describe the emotions in a scene. This experiment also revealed users typically use simple words such as "happy" and "fear" and augment them with more detailed and descriptive words such as "atmospheric" and "claustrophobic". For example, for the movie "Titanic", at the end of the movie where people are drowning and the two main characters were professing their love for each other, an example of the words used for the scene would be "melancholy", "gloomy", "tragic", "depressed", and "true love". In this case only two words "love" and "depressed" are present in the Russell list.

Table 7-10 and Figure 7-4 shows that there was a large gap between the words or labels used by the system and those chosen by the users. Over the course of the affect-based indexing and retrieval system's development, the Russell list was deliberately chosen for its extensive list of words or labels that was designed to cover a wide range of emotions. However, the results presented here illustrate that for an ad hoc IR task where users are able to express their interpretation of the affective features of a video, this is still insufficient. This represents a big semantic gap because as much as 75% of the information supplied by the users have to be ignored if no open-vocabulary query mapping component exists in the system. In an IR task, when the textual descriptions supplied by the users are treated as a query, this means that a large portion of the query supplied by the user would have to be ignored or discarded. In order to achieve this, the open-vocabulary query mapping component of the system presented in Section 6.9 (Open-Vocabulary Query Mapping) is incorporated into the affect-based retrieval system to ensure that this information is not wasted and is incorporated into the query. These

results justify the need for a novel open vocabulary query mapping component as presented in this thesis to be incorporated into the system. Such a component allows any non-expert, English-speaking user to naturally and easily generate queries without having to learn or remember the set of labels that exist in the system. In addition, Table 7-10 quantifies how much of a semantic gap there is between the user and the system.

The three lists scores, shown in Figure 7-5, 7-6, and 7-7 reveal some subtle variations in the performance of the labeling of the system. In Figure 7-5, the Russell list of labels received the highest scores as the words were more detailed and were able to describe the clips more accurately. When the labels were then summarized to Surrey's list of 22 general labels, there was a very slight drop in precision, but the plot of the scores in Figure 7-6 reveals a line concentrated on the range of 4 to 8. This suggested that users were still giving high precision scores, but on the occasions the labels were wrong, the penalty was quite severe, which can be seen in the dips in the plot. There is a further drop in precision when the labels were summarized into Ekman's list of 6 labels. The plot of scores shown in Figure 7-7 shows a line that is concentrated on the range of 1 to 5, with high peaks in between. This suggests that users for the most part found the most general set of labels to be less accurate than the more detailed descriptions. However, when the users found the labels to be correct, it was accurate, as reflected in the high peaks throughout the plot.

7.6.5 Task 3

In the third and final task, for each of the 5 clips, the textual descriptions written down by the user were fed into the system as a query in order to locate relevant clips from the database of the 39 films which were similar in terms of emotional content. Five results clips were retrieved, in order of relevance from top to bottom. Users were then required to watch or browse through the top 5 ranked result clips for each query and rate how similar the result clips are to the original clip for which the query description was written on a scale of 0 to 10 (video similarity score). The 3 lists of detected emotional labels were shown as well, and the users were asked to rate how similar these lists of detected labels

are to the original clip on a scale of 0 to 10 to get a Russell list similarity score, Surrey list similarity score and Ekman list similarity score.

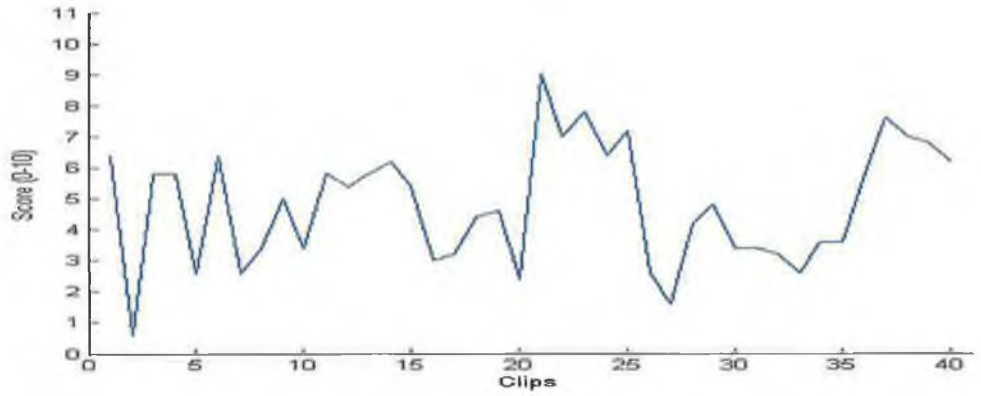


Figure 7-8. Average scores for video similarity for each clip.

Figure 7-8 shows the average scores for clip similarity for each of the 40 original clips. As can be observed, the scores were concentrated between the range 2 and 8, with a mean of 4.8.

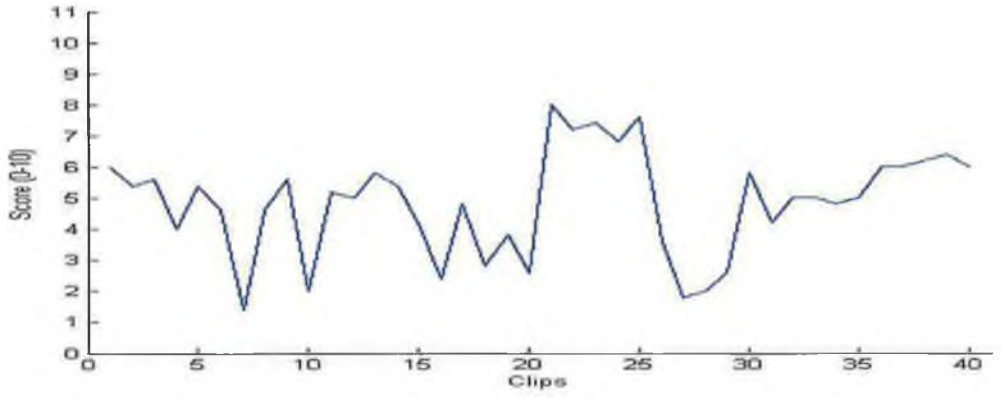


Figure 7-9. Average scores for similarity of results clip's Russell list.

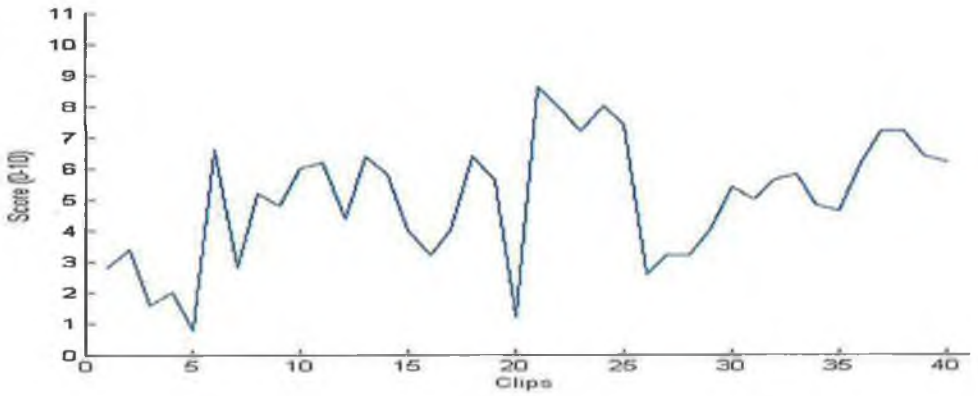


Figure 7-10. Average scores for similarity of results clip's Surrey list.

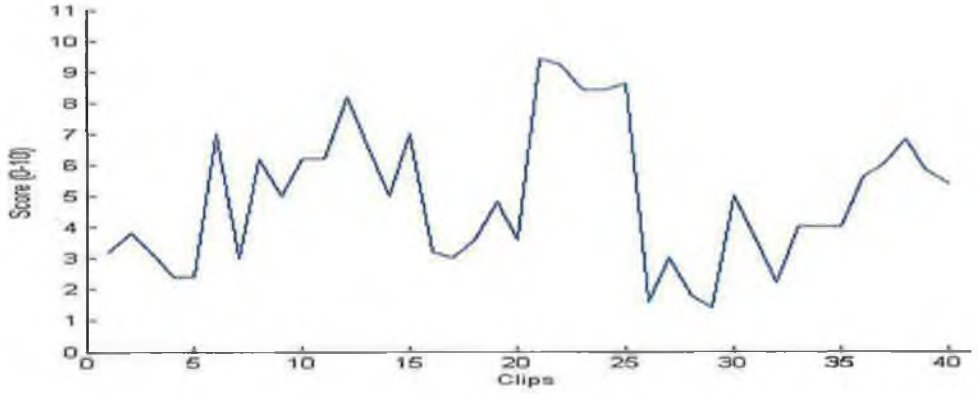


Figure 7-11. Average scores for similarity of results clip's Ekman list.

Therefore Figures 7-9, 7-10 and 7-11 shows the average similarity scores for each of the 3 lists of assigned emotional labels for each of the 40 original clips. The similarity scores compare whether the retrieval result clip are similar to the original clip.

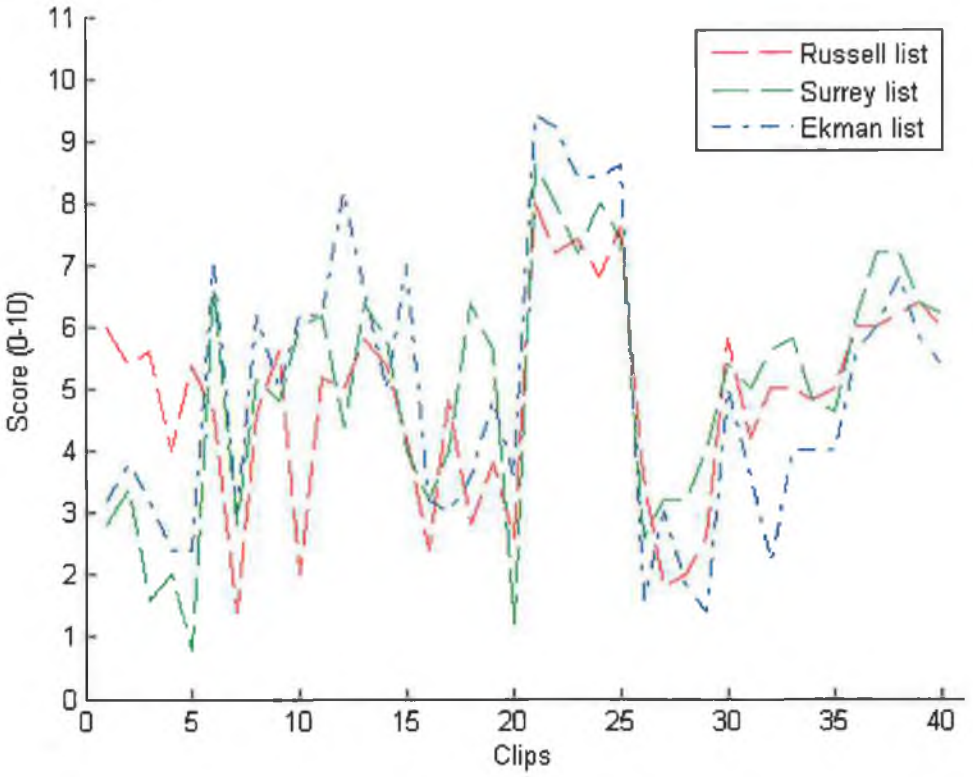


Figure 7-12. Average scores for similarity of result clip's 3 list.

Figure 7-12 shows the average similarity scores for the 3 lists plotted together in the same graph for comparison. The scores are generally in the range 2 to 8, concentrated around 5. This is because the mean values for the Russell, Surrey, and Ekman lists are 4.85, 5, and 5 respectively as shown in Table 7-12. The mean values are calculated by taking the mean values of the similarity scores for the result clips. Therefore the global average score in this case shows the performance of the system at retrieving clips with similar emotional content (relevant clips), giving the system a precision of 0.489.

Table 7-12. Mean values of similarity scores for the results clips.

Types of user scores	**Mean values of scores**
Video similarity	4.80
Russell list similarity	4.85
Surrey list similarity	5.00
Ekman list similarity	5.00
Global average score (Precision)	**0.489**

At the end of the experiment, users were required to answer this question:

"Would you like to see the methods of presenting emotions in this study being used as a tool for searching videos in a video search engine one day? Why?"

7 out of 8 users gave a response of "yes". The remaining user gave a response of "not quite sure" although the user explained that it is useful on a general movie search instead of a specific scene selector. Of the users that gave a response of "yes", two users commented that revealing the precision values for the system would be helpful for them to gauge how well the system is working. Two different users also gave comments that visual representations of emotional content could be incorporated into the system to help users in searching as a future enhancement. The remaining users gave comments that such methods will be useful especially for browsing through movies to look for desired scenes based on moods or emotions.

In addition to answering the question, the users were asked to give a general comment of the experiment, system, and the concept of searching based on affective features. Users commented that the idea of being able to use an affect-based indexing and retrieval system as interesting, innovative, and worth pursuing as further research work. However, they also commented that such research is still in its infancy, requiring more work before it could be deployed for applications in the real-world. There were also comments that it can be difficult sometimes to separate the dominant emotions from the minor ones in a scene.

7.6.6 Discussion of the results of Task 3

Task 3 of the experimental investigation explored the performance of the affect-based indexing and retrieval system. This is necessary to determine the system's performance in retrieving clips with similar emotional content or relevant clips based on an open-vocabulary user query. In order to determine the true performance of the system and reduce bias that might be present on a small result set, a suitably large amount of results

had to be gathered. Each user reviewed 25 results clips, giving a total of 200 clips or 16 hours of results.

The users were required to not only compare the result clips by watching them, but also to compare the result clip's detected emotional labels against the original clip. This was to explore if there were large variations when comparing the clips by just watching the clips or comparing the clips using the detected labels from the 3 lists. A large variance between these scores for the same clip could indicate that the system's result clips and their detected labels were inconsistent. Therefore this extra set of similarity scores from the detected labels are meant to further reinforce the evidence that the system was not only detecting consistent labels but also retrieving consistently relevant clips. As shown in Figure 7-12, most of the scores are very close to each other for each clip. Table 7-12's mean values of the scores show very little difference between each other as well, therefore suggesting that the system was retrieving consistently relevant clips.

It can be observed that the scores are different from user to user. However in each case the user's score were quite consistent in that their scores were always within a small range. For example, user 5 gave very high scores ranging from 7 to 10, which can be seen in the scores between clips 21 to 25 in Figure 7-12. User 6 gave lower scores ranging from 1 to 3, which can be seen between clips 26 to 30. Informal feedback from users reveals that it is possible that different users have different expectations and interpretations of the scale, therefore leading some to be more lenient or conservative in giving out scores. This is why a large number of results were gathered from multiple users. Informal feedback from users also revealed that most of them felt that the system was generally successful in retrieving clips with similar emotions.

Although not directly comparable, the precision result of 0.489 obtained by this system for locating clips with similar emotions holds a lot of promise when compared to other high-level feature detection tasks such as those in [TRECVID].

7.6.7 Known-Item Search Task

In addition to retrieving clips with similar emotional content, a known-item search task was performed after the user experiments had ended to measure the system's performance at retrieving and locating the original film clip described by the text query from the database of films using the textual descriptions of that particular clip's emotional content supplied by the users from task 2.

The textual description for a clip was treated as the user query for that particular clip. Depending on the query processing strategy used, the user query is mapped to the system's Russell list of 151 labels, and these labels used as the final query. The system generates a ranked list of clips from the database of 39 movies using the Okapi BM25 model. The ranked list sorts the clips from top to bottom in order of relevance. From this ranked list, the original clip's position on this list is identified. There were a total of 939 clips from the 39 films therefore the lower its position or number, (towards position 1 in the list) the more accurate the system is at retrieving the original clip based on a user-supplied query. For this task, the objective was to explore the effectiveness of text IR using the affect-based retrieval system. Five different query processing strategies were used to explore and fine tune the system for best retrieval performance as discussed previously in Section 6.9 (Open-Vocabulary Query Mapping). The strategies were "Bypass", "Unweighted full expansion", "Unweighted best expansion", "Weighted full expansion", and "Weighted best expansion".

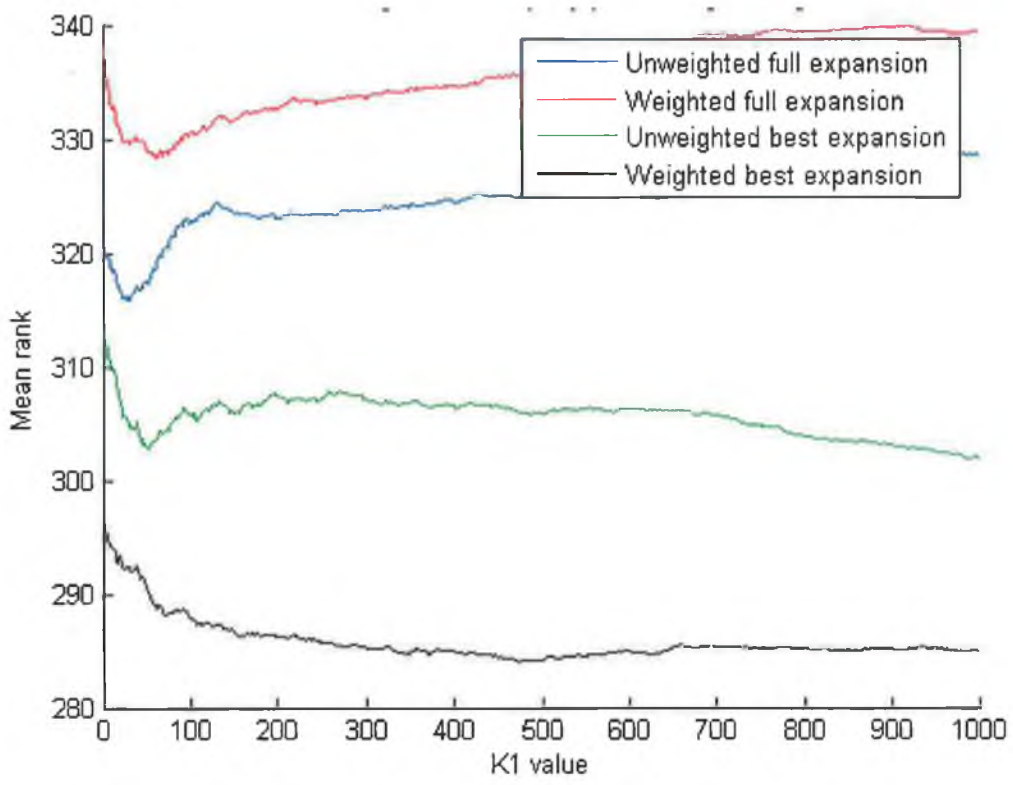


Figure 7-13. Mean rankings for the 4 query processing strategies for different K1 values.

Figure 7-13 shows how different K1 values of the Okapi BM25 IR retrieval model of the IR component of the system affects the mean rankings for the first 4 query processing strategies. A thousand runs were performed using the system automatically to calculate the mean rank for the 40 queries. The b value is set to a constant 0. It can be observed that the "Weighted full expansion" strategy gave the worst mean ranks of 335. The "Unweighted full expansion" strategy gave better mean ranks of 325. This is then followed by the "Unweighted best expansion" strategy with mean ranks of 305. The "Weighted best expansion" strategy retrieved the best mean ranks of 388. There is a noticeable dip for all strategies' mean ranks when the K1 value reaches 51. However, when observed over more K1 values, it is revealed that the best mean rank achieved by the "Weighted best expansion" strategy is when K1 is at the value 474.

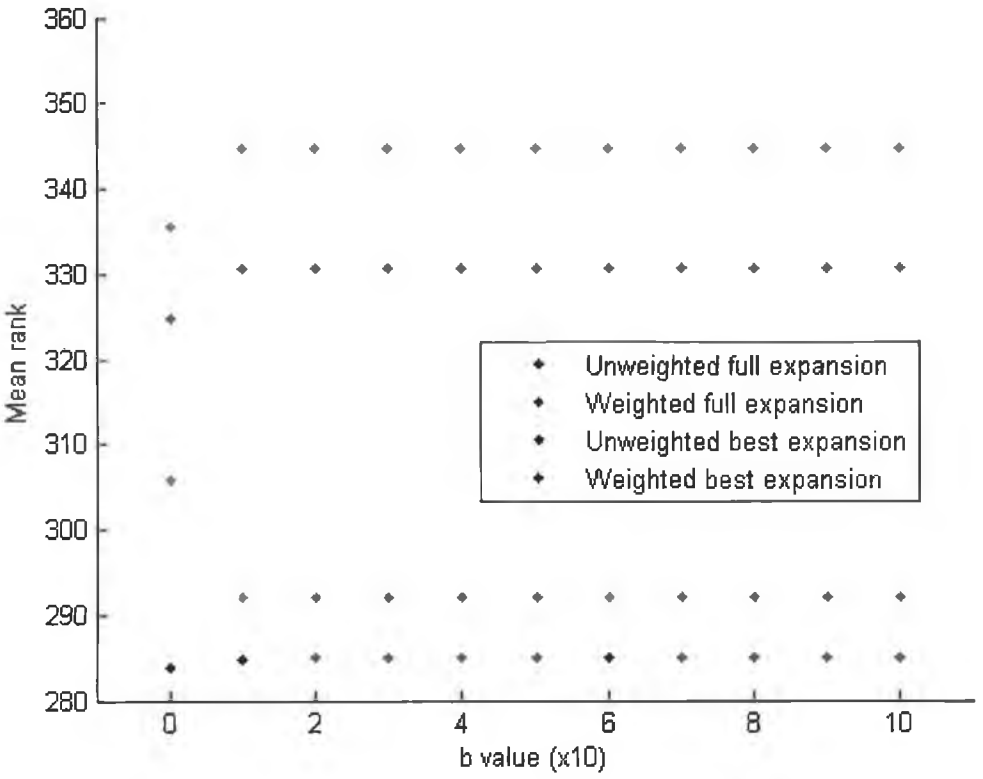


Figure 7-14. Mean rankings for the 4 query processing strategies for different b values.

Figure 7-14 shows how different b values of the IR component of the system affect the mean rankings for the first 4 query processing strategies. The K1 value was set to 474. It can be observed that the best mean ranks were retrieved when the b value is set to 0, except for the "Unweighted best expansion" strategy which retrieved a worse mean rank. When the b values changes from 0 to 1, the mean ranks does not appear to change. Closer inspection of the values reveal that the mean ranks do change with different b values, but the difference was too small to be noticed.

Table 7-13. Ranking results for the 4 query processing strategies (K1 = 474 and b = 0).

Query number:	Unweighted full expansion	Weighted full expansion	Unweighted best expansion	Weighted best expansion
1	**483**	**473**	**371**	**231**
2	0	0	0	0
3	247	260	0	0
4	**346**	**451**	**346**	**451**
5	311	323	0	0
6	**202**	**345**	**223**	**152**
7	**556**	**552**	**523**	**507**
8	**531**	**320**	**341**	**402**
9	**564**	**469**	**396**	**232**
10	**6**	**11**	**7**	**4**
11	0	0	0	0
12	0	0	0	0
13	**35**	**462**	**914**	**931**
14	**301**	**211**	**182**	**94**
15	669	673	0	0
16	**7**	**7**	**7**	**7**
17	**426**	**389**	**388**	**486**
18	0	0	0	0
19	**3**	**4**	**5**	**61**
20	**441**	**441**	**441**	**441**
21	**44**	**44**	**44**	**44**
22	**355**	**355**	**355**	**355**
23	**247**	**323**	**325**	**286**
24	**240**	**240**	**240**	**240**
25	**402**	**350**	**346**	**294**
26	**413**	**380**	**364**	**299**
27	601	636	0	0
28	0	0	0	0
29	**329**	**329**	**329**	**329**
30	0	0	0	0
31	**306**	**304**	**299**	**295**
32	0	0	0	0
33	**414**	**407**	**381**	**398**
34	**445**	**390**	**384**	**217**
35	285	550	0	0
36	**51**	**13**	**12**	**14**
37	0	0	0	0
38	**405**	**336**	**346**	**176**
39	**423**	**378**	**383**	**437**
40	306	311	0	0
Average mean rankings for only values highlighted in bold	**306.73**	**307.07**	**305.84**	**283.96**

Table 7-13 shows the retrieved ranks of the 40 queries for the first 4 query processing strategies. When the system was unable to retrieve the clip, a rank of 0 was returned. The average mean rankings for the values highlighted in bold shows the

retrieval results when the system was able to retrieve the clip using all the four strategies. The "Weighted best expansion" strategy had the best average mean ranks, followed by the "Unweighted best expansion" strategy, "Unweighted full expansion" strategy, and "Weighted full expansion" strategy.

Table 7-14. Number of returned results for the 5 query processing strategies.

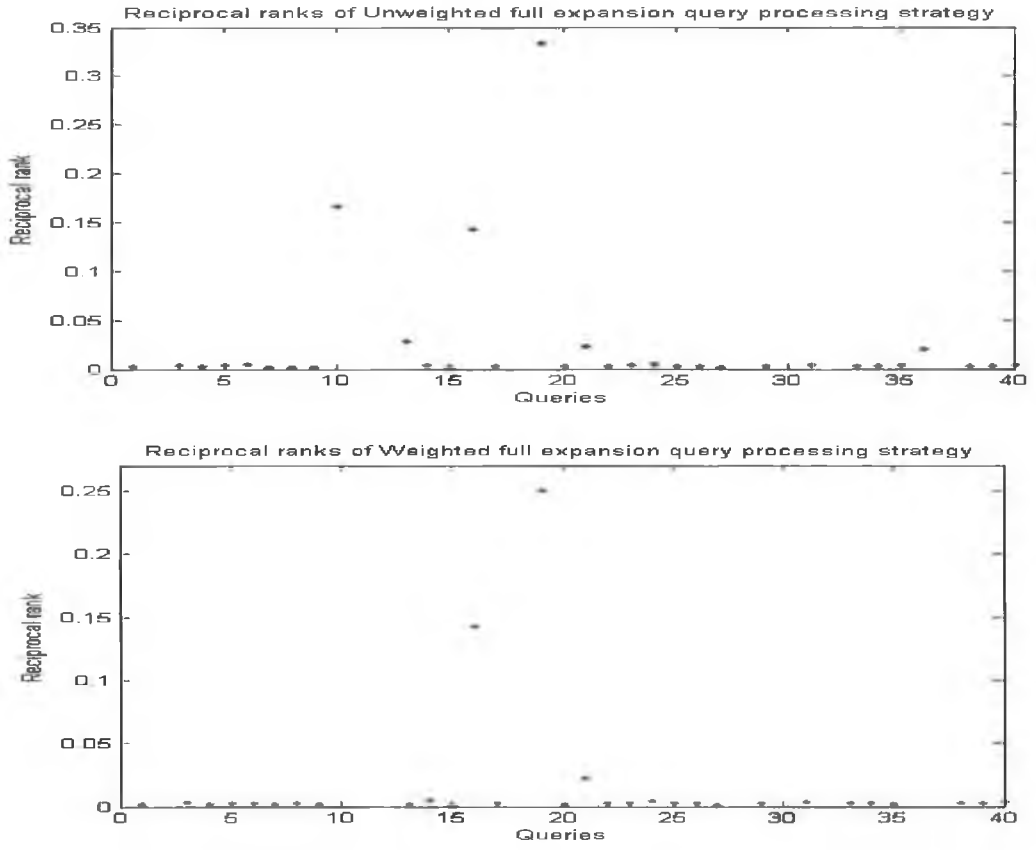
	Unweighted full expansion	**Weighted full expansion**	**Unweighted best expansion**	**Weighted best expansion**	**Bypass**
Number of returned results (out of 40 queries)	32	32	26	26	3
Recall	**0.8**	**0.8**	**0.65**	**0.65**	**0.08**

Table 7-15. Comparison of the ranks for the returned results of the Bypass strategy and the other 4 query processing strategies.

Queries that returned a result for "Bypass"	**Unweighted full expansion**	**Weighted full expansion**	**Unweighted best expansion**	**Weighted best expansion**	**Bypass**
Query 4	346	451	346	451	**58**
Query 14	301	211	182	94	**6**
Query 36	51	13	12	14	**7**

Table 7-14 shows the number of queries for which the system successfully retrieved the target clip using the 5 query processing strategies. Recall can be calculated by dividing the number of returned results by the total number of known results (number of queries). The two "full expansion" strategies managed to retrieve 32 clips out of 40, giving the best recall rate of 0.8. This is followed by the two "best expansion" strategies that managed to retrieve 26 clips with a recall rate of 0.64. The "Bypass" strategy only managed to retrieve 3 clips, with the worst recall rate of 0.08. This indicates that as the user word query is mapped to an increasingly smaller number of labels to form the final query, the recall rate drops. Note that the failure to retrieve a clip at any rank indicates that none of the query words were contained in the label from the Russell list automatically assigned to the target clip in the analysis stage.

Table 7-15 shows the same results for the three queries retrieved by the Bypass strategy and their ranks across different strategies. It can be observed that the "Bypass" strategy achieved the best ranks (highlighted in bold) for all three queries. In addition, the ranks get increasingly worse for the "best expansion" and "full expansion" strategies. This indicates that as the recall rate drops, the system performs better by retrieving clips with ranks that were nearer to 1. This shows that exact matches with the Russell list in the query can be very effective for retrieval, the difficulty with limiting the query vocabulary to only these words is that users are likely to find their list of words constraining and difficult to use in describing their information needs.



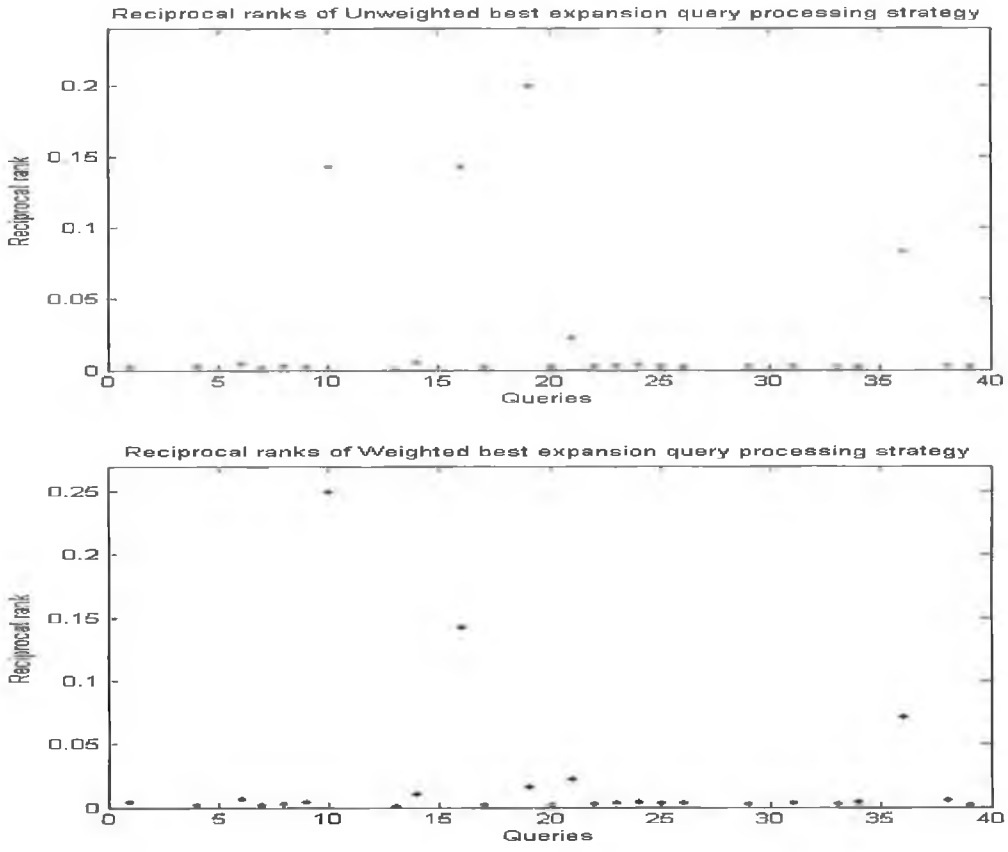


Figure 7-15. Reciprocal ranks for the 4 query processing strategies.

Table 7-16. The mean-reciprocal-rank (MRR) for the 4 query processing strategies.

	Unweighted full expansion	Weighted full expansion	Unweighted best expansion	Weighted best expansion
MRR	0.024	0.020	0.025	0.022

Figure 7-15 shows the reciprocal ranks of the 40 queries that were processed using the first 4 query processing strategies. Table 7-16 shows the calculated mean-reciprocal-rank (MRR) of the 4 strategies. It can be observed that the "Unweighted best expansion" strategy achieved the best MRR of 0.025, followed by the "Unweighted full expansion" strategy with a MRR of 0.024, "Weighted best expansion" strategy with a MRR of 0.022, and "Weighted full expansion" strategy with a MRR of 0.020. When compared with previous results in Table 7-13, the results show that the "Unweighted" strategies returned very good ranks for a few queries compared to the "Weighted" strategies. However, on average the "Weighted" strategies retrieved better ranks when more queries are considered.

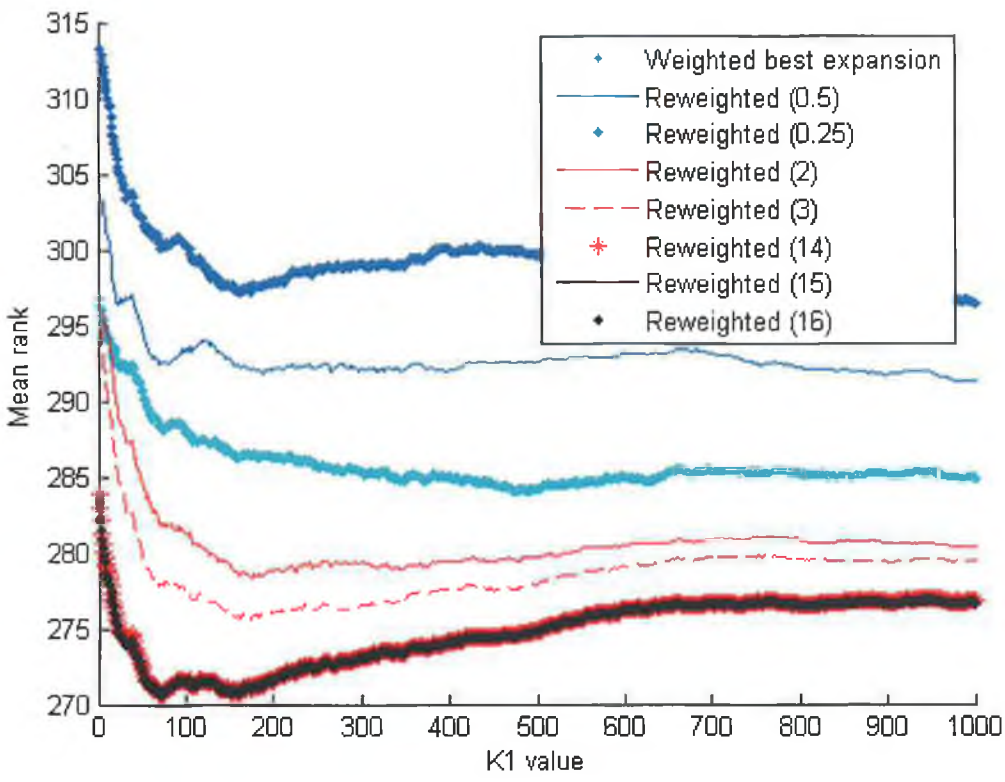


Figure 7-16. Mean rankings for different reweighted and K1 values of "Weighted best expansion" strategy.

Using the "Weighted best expansion" strategy which retrieved the best ranks, the relatedness scores were re-weighted to explore if improvements could be made to the mean ranks. The relatedness scores were first raised by an exponent value before it was multiplied with the combined weights of the Okapi BM25 IR model. Figure 7-16 shows the mean ranks for the re-weighted strategy, where the numerical value enclosed in round brackets are the exponent value. It can be observed that as the exponent value is decreased (0.5 and 0.25), the mean ranks get worse. However, when the exponent value is increased, the mean ranks improve over the original "Weighted best expansion" strategy (plotted in a light blue colour). The mean ranks improvement stops when the exponent value reaches 14, as can it be seen that the black plot line (15 and 16) rests on top of the red star line (14). In addition, the re-weighted strategies were processed with a

b value of 0 and 1000 different K1 values, similar to Figure 7-13. This revealed that the best K1 value for the re-weighted strategies were 71.

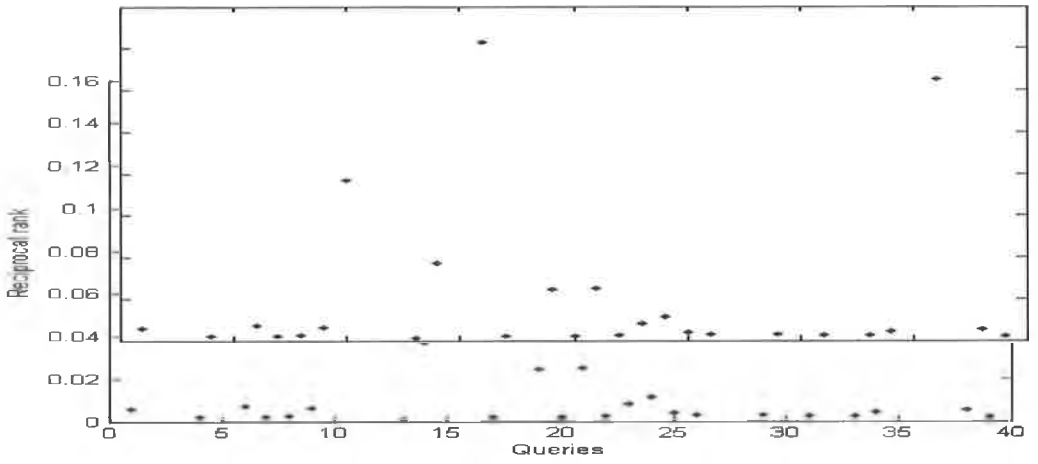


Figure 7-17. Reciprocal ranks of Reweighted best expansion (16) query processing strategy.

Table 7-17. The mean-reciprocal-rank (MRR) of the re-weighted strategy.

	Reweighted best expansion (16)
MRR	0.020

Figure 7-17 show the reciprocal ranks of the re-weighted best expansion strategy with an exponent value of 16. Table 7-17 show the MRR value of the re-weighted best expansion strategy. When compared with Figure 7-15, it can be observed that the re-weighted strategy retrieved better ranks on average, but did not improve on the previous MRR value of 0.025.

7.6.8 Discussion on the Results of the Known-Item Search Task

The known-item search task was aimed at fine-tuning and improving the system's retrieval performance. Five different query processing strategies were investigated using different values for the K1 and b tuning constant of the Okapi BM25 IR model of the IR component of the system. The improvement in results was shown by plotting the mean ranks of the retrieved clips.

As was observed in Figure 7-13, the differences between the four different strategies' mean ranks were clear. The strategies that mapped the user query words to the most number of labels (full expansion) had the worst ranks. When the mapping used was more restricted (best expansion), the system retrieved better ranks. The "weighting" strategies also had a considerable impact on the ranks. It is observed that the weighting retrieves better ranks for the best expansion strategy but retrieves worse ranks for the full expansion strategy. In addition, using exponential values for the re-weighted best expansion strategy further improves the mean rankings as shown in Figure 7-16.

The K1 and b tuning constant is usually determined empirically in standard text IR systems. Therefore an experimental investigation was conducted using the four different query processing strategies to determine the best value to set for the two tuning constant. The best K1 and b values results retrieve the best ranks for the IR system. It was found from Figure 7-13 that 51 was a good value for K1 as there was a dip in all four strategies' mean ranks. However, the K1 value of 474 was determined to be the best value as it retrieves the lowest mean ranks for the "Weighted best expansion" strategy. When the "Weighted best expansion" strategy was re-weighted, it was found that a K1 value of 71 retrieved the best ranks on average. The best b value of 0 was determined from Figure 7-14 as there was no more noticeable improvement in mean ranks after the b value changes to 0.1 and above.

On average the best mean rankings were derived from the weighted and re-weighted strategies, however the MRR values from Table 7-16 and 7-17 show that the unweighted strategies returned the best ranks for a select few queries.

Tables 7-13, 7-14 and 7-15 show that there was a slight tradeoff between recall and mean ranks. When the recall rates increase, the mean ranks were worse, and when the recall decreases the mean ranks were better.

The results of this task shows clips which had been described by users using text queries can be retrieved with a measure of consistency. This also indicates that the open-vocabulary query mapping can be improved with different strategies and fine tuning. The best mean ranks of around 280 out of the worst possible 939 (total number of clips in the database) shows that the system was successful in indexing, sorting, and retrieving relevant clips, although due to the relatively poor rank of the desired known-item, it is not

clear to what extent this could be used as a useful known-item retrieval system in a real user task.

7.7 Summary

The three main questions that were set out at the beginning of this chapter have been answered by the results of the experiments. In Task 1, the multiple human-verified results of the valence and arousal curves show that the system is capable of detecting affective features. The first attempt at quantifying the valence and arousal modeling in this thesis also revealed an optimistic analysis that that valence, while harder to detect or describe than arousal, can be detected with similar reliability to arousal.

Task 2 shows that the open-vocabulary query mapping techniques presented in this thesis are vital towards retrieving relevant clips. The experiment also revealed the wide range of words that users choose to describe emotional content in a film clip.

Task 3 and the Known-Item Search Task reveal that the system is able to retrieve relevant clips and the relatively small variance in the similarity scores indicate that the system was retrieving consistently relevant clips.

The experimental investigation presented in this chapter was quite open-ended as it was difficult to predict how such a high-level affect-based system would perform. There was no similar previous experimental work to be found in published literature, therefore there were no baseline results and methodology to use as a guide. The methodology and experimental work presented here therefore can be used as a guide and reference for future work in affect-based retrieval.

Chapter 8 Conclusions and Future Work

This thesis has presented a novel affect-based indexing and retrieval system for films that combines techniques from the fields of signal processing, affective computing, digital video and text retrieval. Emotional content in videos can offer a new, natural and intuitive dimension of interaction for users in information retrieval tasks as well as narrowing the semantic gap between user queries and system-detected features in multimedia content.

Research into human cognitive psychology reveals that emotions play an integral part in human decision-making and interaction. A human's emotional state can be difficult to detect without the use of intrusive physiological sensors, however analysis of affective features in videos offers an alternative method of detecting emotions. Although human viewers can have varying subjective interpretations and experience different emotional states when watching a video, the affective features in videos do not change. This provides consistency for the subjective task of indexing and retrieving videos based on emotional content. Therefore, this thesis presents the justifications, a review of relevant state-of-the-art research, and the design and experimental investigation of a prototype affect-based multimedia retrieval system.

The affect-based indexing and retrieval system presented in this thesis combine research reported in [Hanjalic and Xu, 2003] and [Kok, 2006] to use low-level audio and visual features to model the affective features. The system automatically extracts low-level audio and visual features that are combined to model affective features. These affective features are then mapped to pre-defined emotional verbal labels that are used to index the videos. Videos with the pre-defined emotional labels can then be retrieved from a video database using text information retrieval techniques. User text queries can include words that do not exist in the system. These words are mapped using the WordNet ontology onto the system's pre-defined labels of the system to retrieve the relevant videos. This provides users with a natural and intuitive method of interacting with the system to search for videos with the desired emotional content.

An experimental investigation was conducted to explore how the system performs. A set of test data in the form of 39 Hollywood movies from a wide range of

genres were used to test the system. The experimental investigation involving human volunteers concluded that the system was able to retrieve relevant videos based on their text queries expressing the desired emotional content. Comments from the volunteers revealed that the system and research work is novel and interesting, and that further research work is worth pursuing for its potential in information retrieval applications.

The affect-based indexing and retrieval system as presented in this thesis has demonstrated retrieval of relevant videos from a database of films. However, the field of such research work is still in its infancy, and it is anticipated the system's results can be further improved by a better understanding of how affective features can be modeled.

Information from high-level feature detectors such as facial gesture detection, music analysis, and object detection could be incorporated into the system. In dramatic conversational scenes, the character's face is often in clear view, therefore a facial gesture detector could be used to recognize what emotions the characters are trying to express. The audio analyses that are used in the affect-based retrieval system presented in this thesis can be further extended as well. A speech/music detector can be used to separate speech and music from the audio track for further processing and analyses. Music plays an important part of setting up the mood of a scene, often before anything happens visually. The emotional content of the music could be used to assign better emotional labels to scenes that do not have much visual activity or information. It might even be possible to use music to adjust or bias the system towards a certain emotion before other less reliable features is detected and used. Object or shape detectors could also be used to detect objects that have pre-defined emotional meaning. For example, scenes with guns, knives and explosions could indicate the scene as being violent, which is in contrast to outdoor scenes of trees, mountains and lush green hills.

Although research work for the feature detectors as described is ongoing, it is not clear how this information can be used or relate to the modeling of the affective features. Future work into better modeling of affective features is vital for the improvement in detecting affective features, and subsequently, emotion. Therefore as future work, the system presented in this thesis could be used to explore and quantify how much improvement in affective retrieval performance can be gained when new affective feature

modeling strategies are introduced into the system by comparing the known-item retrieval rankings.

Recent research work in detecting affective features is reliant upon common conventions or knowledge of film-making or "film grammar". The low-level features used for modeling the affective features make the assumption that the "rules" or guidelines of film grammar are always being followed. Therefore as future work, the system could be used to explore the absence or violation of film grammar by looking at areas where the system performs poorly.

Another interesting area of future work is to look at the system-detected emotional labels and compare them with different film genres. Clusters of emotional labels and different relationships between them might be discovered for different film genres that could be used to classify the films. However, such an application would require a far larger number of movies than were available for the experiments for this thesis. Therefore with a larger set of movies such an application could be investigated for future work.

The affect-based system can also be incorporated into existing video retrieval systems to explore the effects of incorporating affective information with other standard video retrieval features. The additional metadata describing affective information could be used to help improve retrieval results by giving users an additional feature to filter out non-relevant videos based on the affective content. The system presented in this thesis therefore not only takes us one step closer to a real-world system for browsing videos based on emotions, but also can be used as a research tool to test and explore new ideas in the field of affective computing and information retrieval.

The results of the known-item affect-based retrieval experiments in this thesis have demonstrated that accuracy of retrieval can be improved by using the full range of the words entered by users. In the experiments, this was achieved by mapping the words outside the system's vocabulary of 151 words using the WordNet ontology. It would be interesting to explore the extent of which performance might be further improved by expanding the vocabulary of the core affect-based retrieval system beyond the "Russell" list of 151 words for which the arousal and valence values are known. This would require a further data collection exercise of the form described by Russell and Mehrabian [1977].

In this exercise, multiple subjects were asked to score the arousal, valence and dominance dimension of the words. For an expanded version, a much larger list of words would need to be scored, thus requiring a larger number of human subjects to provide the ratings of the words.

In addition, the experimental investigation has shown that performance with the open-vocabulary system varied when different criteria were applied for term selection from and alternative word similarity measures between the query word and the system affect vocabulary were used. It was shown that using a weighting function can be beneficial and that performance can be further improved by modifying the dynamic range of this associated score. The Hirst and St-Onge relatedness measure was used in this experiment because it is an established association scoring mechanism in WordNet. Based on this metric, the experiments reported in this thesis provide a good baseline, however it would be interesting to consider whether there might be alternative word association scoring metrics which might be effective in this application.

References

[Bach et al., 1996] Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., Shu, C.: Virage image search engine: An open framework for image management, *Proceedings of the SPIE Storage and Retrieval for Still Image and Video Databases*, (1996), 76-87.

[Barjatya, 2004] Barjatya, A.: Block Matching Algorithms for Motion Estimation, *Digital Image Processing 6620 Final Project Paper*, Utah State University, (2004).

[Brown et al., 1994] Brown, M.G., Foote, J.T., Jones, G.J.F., Jones, K.S., Young, S.J.: Video Mail Retrieval using Voice: An overview of the Cambridge/Olivetti retrieval system, *Proceedings of ACM Multimedia 94 Workshop on Multimedia Database Management Systems*, San Francisco, USA, (1994), 47-55.

[Brown et al., 1995] Brown, M.G., Foote, J.T., Jones, G.J.F., Jones K.S., Young, S.J.: Automatic Content-Based Retrieval of Broadcast News, *Proceedings of ACM Multimedia 95*, San Francisco, USA, (1995), 35-43, 1084-1090.

[Browne and Smeaton, 2004] Browne, P., Smeaton, A.F.: Video Information Retrieval Using Objects and Ostensive Relevance Feedback, *ACM Symposium on Applied Computing*, Nicosia, Cyprus, (SAC 2004).

[Browne et al., 2002] Browne, P., Smeaton, A.F., Murphy, N., O'Connor, N., Marlow, S., Berrut, C.: Evaluating and Combining Digital Video Shot Boundary Detection Algorithms, *Proceedings of Irish Machine Vision and Image Processing Conference*, (2002).

[CDVP] *Centre for Digital Video Processing Website*, Available online at URL: http://www.cdvp.dcu.ie (last visited November 2006)

[Christel and Conescu, 2005] Christel, M., Conescu, R.: Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries, *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital Libraries*, Denver, USA, (2005), 69-78.

[Cooley and Tukey] Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complext Fourier series, *Math. Computation*, (1965), 19:297-301.

[Collet et al., 1997] Collet, C., Vernet-Maury, E., Delhomme, G., Ditmar, A.: Autonomic Nervous System Response Patterns Specificity to Basic Emotions, *Journal of the Autonomic Nervous System*, 62(1-2), (1997), 45-57.

[Cornelius, 1996] Cornelius, R.: *The Science of Emotion*, Prentice Hall, New Jersey, USA, (1996).

[Cytowic, 1989] Cytowic, R.E.: *Synesthesia: a union of the senses*, Springer-Verlag, New York, USA, (1989).

[Czirjek et al., 2003] Czirjek, C., O'Connor, N., Marlow, S., Murphy, N.: Face Detection and Clustering for Video Indexing Applications, *Proceedings of Advanced Concepts for Intelligent Vision Systems*, Ghent, Belgium, (Acivs 2003).

[Damasio, 1994] Damasio, A.R.: Descartes' Error: *Emotion, Reason and the Human Brain*, Gosset/Putnam Press, New York, USA, (1994).

[Darwin, 1872] Darwin, C.: *The Expression of Emotions in Man and Animals*, University Chicago Press, Chicago, IL, USA, (1872/1965).

[Derry, 1999] Derry, S.J.: *A Fish called Peer Learning: Searching for Common Themes*, A.M. O'Donnell & A. King, (1999).

[Detenber et al., 1997] Detenber, B.H., Simons, R.F., Bennett, G.G.: Roll 'em!: The effects of picture motion on emotional responses", *Journal of Broadcasting and Electronic Media*, 21, (1997), 112-126.

[Dietz and Lang, 1999] Dietz, R., Lang, A.: *Aefective Agents: Effects of Agent Affect on Arousal, Attention, Liking and Learning*, http://www.polara.org/pubs/cogtech1999.html.

[DivX] *DivX Video Codec Website*, Available online at URL: http://www.divx.com (last visited November 2006)

[Eakins, 1996] Eakins, J.P.: Automatic image content retrieval – are we getting anywhere?, *Proceedings of the 3rd International Conference on Electronic Library and Visual Information Research*, (1996), 123-135.

[Ekman and Friesen, 1978] Ekman, P., Friesen, W.V.: *Facial Action Coding System*, Consulting Psychologists Press Inc., 577 College Avenue, Palo Alto, California 94306, USA, (1978).

[Ekman et al., 1987] Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., Scherer, K.R., Tomita, M., Tzavaras, A.: Universals and Cultural Differences in the Judgements of Facial Expressions of Emotion, *Journal of Personality and Social Psychology*, 53, (1987), 712-717.

[Essa and Pentland, 1997] Essa, I., Pentland, A.: Coding, analysis, interpretation and recognition of facial expressions, *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, (1997), 757-763.

[Flickner et al., 1995] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Qian, H.D.B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The QBIC system, *IEEE Computing*, (Sept 1999), 23-32.

[Gemmell et al., 2004] Gemmell, J., Williams, L., Wood, K., Lueder, R., and Bell, G.: Passive Capture and Ensuing Issues for a Personal Lifetime Store, *The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, Columbia University, New York (CARPE 2004), 48-55.

[Gilvarry, 1999] Gilvarry, J.: Calculation of Motion Using Motion Vectors Extracted from an MPEG Stream, *Technical Report, School of Computer Applications*, Dublin City University, (1999).

[Google Video] *Google Video Website*, Available online at URL: http://video.google.com (last visited November 2006)

[Hang, 2003] Hang, B.K.: Affective Content Detection using HMMs, *Proceedings of the 11th ACM International Conference on MultiMedia 2003*, Berkeley, California, USA, (MM'03), 259-262.

[Hanjalic and Xu, 2003] Hanjalic, A., and Qun Xu, L.: User-oriented Affective Video Content Analysis, *Proceedings of the 36th Annual Simulation Symposium* (ANSS'03), IEEE (2003).

[Harman, 1997] Harman, D.K.: *The Fifth Text REtrieval Conference (TREC-5)*, National Institute of Standards and Technology, Gaithersburg, MD, (1997).

[Hauptmann, 2005] Hauptmann, A.G.: Lessons for the Future from a Decade of Informedia Video Analysis Research, *Proceedings of the 4th International Conference on Image and Video Retrieval*, Singapore, (CIVR 2005), 1-10.

[Hauptmann and Christel, 2004] Hauptmann, A.G., Christel, M.G.: Successful Approaches in the TREC video retrieval Evaluations, *Proceedings of ACM Multimedia 2004*, New York, ACM (2004) 668-675.

[Haykin, 1991] Haykin, S.: *Adaptive Filter Theory*, Prentice Hall, (1991).

[Hirst and St-Onge, 1998] Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms, *Wordnet: An Electronic Lexical Database*, MIT Press, (1998), 305-332.

[Inanoglu and Caneel, 2005] Inanoglu, Z., Caneel, R.: Emotive Alert: HMM-Based Emotion Detection in Voicemail Messages, *Proceedings of the International Conference on Intelligent User Interfaces 2005*, San Diego, California, USA, (IUI 2005), 251-253.

[Informedia] *Informedia Website*, Available online at URL: http://www.informedia.cs.cmu.edu (last visited November 2006)

[Izard, 1993] Izard, C.E.: Four systems for emotion activation: Cognitive and non-cognitive processes, *Psychological Review*, 100(1), (1993), 68-90.

[James, 1890] James, W.: *The Principles of Psychology*, Dover Publications, (1890).

[Jarina et al., 2001] Jarina, R., Murphy, N., O'Connor, N., Marlow, S.: Speech-Music Discrimination from MPEG-1 Bitstream, Kluev V.V and Mastorakis N.E. (Eds.), *Proceedings of Advances in Signal Processing, Robotics and Communications*, WSES Press, (2001), 174-178.

[Kagan, 1984] Kagan, J.: *The Nature of the Child*, Basic Books, New York, USA, (1984).

[Kantor and Voorhees, 1999] Kantor, P.B., Voorhees, E.M.: *The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text, Information Retrieval, 2(2/3)*, Kluwer Academic Publishers, (2000), 165-176.

[Kok, 2006] de Kok, I.: A Model for Valence Using a Color Component in Affective Video Content Analysis, *Proceedings of the 4th Twente Student Conference on IT, Faculty of Electrical Engineering, Mathematics and Computer Science*, University of Twente, (2006).

[Lanzetta and Orr, 1986] Lanzetta, J.T., Orr, S.P.: Excitatory Strength of Expressive Faces – Effects of Happy and Fear Expressions and Context on the Extinction of a Conditioned Fear Response, *Journal of Personality and Social Psychology*, 50(1), (1986), 190-194.

[Leacock and Chodorow, 1998] Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification, *WordNet: An electronic lexical database*, MIT Press, 265-283.

[Lee et al., 1999] Lee, H., Smeaton, A.F., Furner, J.: User-interface Issues for Browsing Digital Video, *Proceedings of the 21st Annual Colloquium on IR Research*, Glasgow, UK, (IRSG'99).

[Lee et al., 2000] Lee, H., Smeaton, A.F., McCann, P., Murphy, N., O'Connor, N., Marlow, S.: Físchlár on a PDA: A Handheld User Interface to a Video Indexing, Browsing and Playback System, *ERCIM Workshop "User Interfaces for All"*, Florence, Italy, (ERCIM 2000).

[Lew et al., 2006] Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based Multimedia Information Retrieval: State of the Art and Challenges, *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1), (2006), 1-19.

[Liu et al., 2003] Liu, H., Selker, T., Lieberman, H.: Visualizing the Affective Structure of a Text Document, *Proceedings of the Conference on Human Factors in Computing Systems*, Ft. Lauderdale, Florida, USA, (CHI'2003), 740-741.

[Matlab] *Matlab Website*, Available online at URL: http://www.mathworks.com/ (last visited November 2006)

[McMahon, 1997] McMahon, M.: Social Constructivism and the World Wide Web – A Paradigm for Learning, *Proceedings of Australasian Society for Computers in Learning In Tertiary Education*, Perth, Australia, (ASCILITE 1997).

[MPEG] *MPEG Website*, Available online at URL: http://www.chiariglione.org/mpeg/ (last visited November 2006)

[Murray and Arnott, 1993] Murray, I.R., Arnott, J.L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *Journal of Acoustic Society of America*, 93(2), (1993), 1097-1108.

[Nass et al., 1994] Nass, C.I., Steuer, J.S., Tauber, E.: Computers are Social Actors, *Proceedings of the Conference on Human Factors in Computing Systems*, Boston, MA, USA, (CHI'1994), 72-78.

[Nasoz et al., 2003] Nasoz, F., Alvarez, K., Lisetti, C.L., Finkelstein, N.: Emotion Recognition from Physiological Signals for Presence Techonologies, International *Journal of Cognition, Technology, and Work – Special Issue on Presence*, 6(1), (2003).

[Nyquist, 1928] Nyquist, H.: Certain topics in telegraph transmission theory, *Transactions of American Institute of Electrical Engineers* (AIEE), 47, (April 1928), 617-644.

[OECD] *Organization for Economic Co-operation and Development Broadband Statistics Website*, Available online at URL: http://www.oecd.org/sti/ict/broadband (last visited November 2006)

[Ortony et al., 1988] Ortony, A., Clore, G.L., Collins, A.: *The Cognitive Structure of Emotions*, Cambridge University Press, (1988).

[Osgood et al., 1957] Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: *The measurement of meaning*, University of Illinois Press, 1975.

[Over et al., 2006] Over, P., Ianeva, T., Kraaij W., Smeaton, A.F.: *TRECVID 2005 – An Overview*, National Institute of Standards, (2006).

[O'Toole et al., 1999] O'Toole, C., Smeaton, A.F., Murphy N., Marlow, S.: Evaluation of Automatic Shot Boundary Detection on a Large Video Test Suite, *The Challenge of Image Retrieval: 2nd UK Conference on Image Retrieval, Newcastle*, UK, (CIR 1999).

[Pan, 1995] Pan, D.: A Tutorial on MPEG/Audio Compression, *Proceedings of IEEE Multimedia 1995*, 2(2), (1995), 60-74.

[Patwardhan et al., 2003] Patwardhan, S., Banerjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation, *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, (CICLing 2003), 241-257.

[Pedersen et al., 2004] Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet::Similarity – Measuring the Relatedness of Concepts, *Proceedings of the 19th National Conference on Artificial Intelligence* (AAAI 2004), San Jose, CA, USA.

[Picard, 1997] Picard, R.W.: *Affective Computing*, MIT Press, Cambridge, MA, USA, (1997).

[Picard, 2003] Picard, R.W.: Affective Computing: Challenges, *International Journal of Human-Computer Studies*, 59(1-2), (2003), 53-64.

[Picard et al., 2001] Picard, R.W., Vyzas, E., Healey, J.: Towards Machine Emotional Intelligence: Analysis of Affective Physiological State, *IEEE Transactions Pattern Analysis and Machine Intelligence*, 23(10), (2001).

[Picard and Cosier, 1997] Picard, R.W., Cosier, G.: Affective Intelligence – the missing link?, *BT Technology Journal*, 14(4), (1997) 150-161.

[Rasheed et al., 2005] Rasheed, Z., Sheikh, Y., Shah, M.: On the Use of Computable Features for Film Classification, *Proceedings of IEEE Transactions on Circuits and Systems for Video Technology*, 15(1), (2005).

[ReplayTV] *ReplayTV Digital Video Recorder Website*, Available online at URL: http://www.replaytv.com

[Resnik, 1995] Resnik, P.: Using information content to evaluate semantic similarity, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, Canada, (1995), 488-453

[Robertson and Sparck Jones, 1997] Robertson, S.E., Sparck Jones, K.: Simple, proven approaches to text retrieval, *Technical Report, TR356*, Cambridge University Computer Laboratory, (1997).

[Rothwell et al., 2006] Rothwell, S., Lehane, B., Chan, C.H., Smeaton, A.F., O'Connor, N.E., Jones, G.J.F., Diamond, D.: The CDVPlex Biometric Cinema: Sensing Physiological Responses to Emotional Stimuli in Film, *Adjunct Proceedings of the 4th International Conference on Pervasive Computing*, Dublin, Ireland, (2006).

[Rui et al., 2000] Rui, Y., Gupta, A., Acero, A.: Automatically Extracting Highlights for TV Baseball Programs, *Proceedings of ACM Multimedia 2000*, Los Angeles, CA USA, (2000), 105-115.

[Russell and Mehrabian, 1977] Russell, J.A., Mehrabian, A.: Evidence for a Three-Factor Theory of Emotions, *Journal of Research in Personality*, 11, (1977), 273-294.

[Sadlier and O'Connor, 2005] Sadlier, D., and O'Connor, N.: Event Detection based on Generic Characteristics of Field Sports. *IEEE International Conference on Multimedia and Expo*, Amsterdam, The Netherlands, (ICME 2005).

[Salton, 1983] Salton, G., McGill, M.: *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[Salway and Graham, 2003] Salway, A., Graham, M.: Extracting Information about Emotions in Films, *Proceedings of the 11th ACM International Conference on Multimedia*, Berkeley, CA, USA, (2003), 299-302.

[Salway et al., 2005] Salway, A., Vassiliou, A., Ahmad, K.: What Happens in Films?, *IEEE International Conference on Multimedia and Expo* (ICME 2005), 49-52.

[SenseWear] *SenseWear armband by Bodymedia Website*, Available online at URL: http://www.bodymedia.com/products/bodymedia.jsp

[Shugrina et al., 2006] Shugrina, M., Betke, M., Collomosse, J.P.: Empathic Painting: Interactive stylization using observed emotional state, *Proceedings of the 4th International Symposium on Non-photorealistic Rendering and Animation*, (NPAR 2006), 87-96.

[Silverstein et al., 1999] Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a Very Large Web Search Engine Query Log, *Proceedings of ACM SIGIR Forum*, 33(1), (1999), 6-12.

[Simons et al., 1999] Simons, R., Detenber, B.H., Roedema, T.M., Reiss, J.E.: Emotion-processing in three systems: The medium and the message, *Psychophysiology*, 36, (1999), 619-627.

[Smeaton et al., 2004] Smeaton, A.F., Lee, H., McDonald, K.: Experiences of creating four video library collections with the Físchlár system, *International Journal on Digital Libraries: Special Issue on Digital Libraries as Experienced by the Editors of the Journal*, 4(1), (August 2004).

[Smeaton, 2004] Smeaton, A.F.: Indexing, Browsing, and Searching of Digital Video, *Annual Review of Information Science and Technology*, 38(8), (2004) 371-407.

[Smeulders et al., 2000] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), (2000), 1349-1380.

[Smith and Kanade, 1995] Smith, M., Kanade, T.: Video Skimming for Quick Browsing Based on Audio and Image Characterization, *Carnegie Mellon University technical report* CMU-CS-95-186, (July 1995), *Pattern Analysis and Machine Intelligence Journal*, (PAMI 1995).

[Slaney and McRoberts, 2003] Slaney, M., McRoberts, G.: BabyEars: A recognition system for affective vocalizations, *Journal of Speech Communications*, 39(3), (2003), 367-384.

[Sparck Jones, 1994] Sparck Jones, K.: Reflections on TREC, *Technical Report 347*, Cambridge University Computer Laboratory, (August 1994).

[TiVo] *TiVo Digital Video Recording Device Website*, Available online at URL: http://www.tivo.com (last visited November 2006)

[TRECVID] *TRECVID Video Retrieval Evaluation Website*, Available online at URL: http://www-nlpir.nist.gov/projects/trecvid/ (last visited November 2006)

[Van Rijsbergen, 1979] Van RijsBergen, C.J.: *Information Retrieval*, 2nd Edition, Butterworths, London, 1979.

[Valdez and Mehrabian, 1994] Valdez, P., Mehrabian, A.: Effects of color on emotions, *Journal of Experimental Psychology: General*, 123(4), (1994), 394-409.

[VLC] *VLC media player Website*, Available online at URL: http://www.videolan.org/vlc/

[Wikipedia] *Wikipedia Website*, Available online at URL: http://www.wikipedia.org/ (last visited November 2006)

[WordNet] *Online lexical reference dictionary WordNet Website*, Available online at URL: http://wordnet.princeton.edu/ (last visited November 2006)

[WordNet::Similarity] *WordNet::Similarity Website,*, Available online at URL: http://sourceforge.net/projects/wn-similarity (last visited November 2006)

[YouTube] *YouTube Website*, Available online at URL: http://www.youtube.com (last visited November 2006)

[Zhai et al., 2004] Zhai, Y., Shah M., Rasheed, Z.: A framework for semantic classification of scenes using finite state machines, *Proceedings of the Conference for Image and Video Retrieval*, (CIVR 2004), 279-288.

[Zhang and Jay Kuo, 2001] Zhang, T., Jay Kuo, C.-C.: *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*, Kluwer Academic Publishers, (2001).

Appendix A: Russell List and Surrey List of Emotional Verbal Labels

Russell list of verbal labels:

Bold	Overwhelmed	Pain
Useful	Curious	Indignant
Mighty	Relaxed	Repentant
Kind	Untroubled	Sinful
Satisfied	Modest	Shy
Admired	Secure	Guilty
Proud	Nonchalant	Weary
Interested	Aloof	Angry
Arrogant	Leisurely	Confused
Inspired	Reserved	Dissatisfied
Excited	Protected	Regretful
Influential	Consoled	Tense
Aggressive	Quiet	Disdainful
Strong	Sheltered	Depressed
Dignified	Humble	Despairing
Powerful	Solemn	Lonely
Elated	Reverent	Meek
Hopeful	Astonished	Burdened
Triumphant	Disgusted	Timid
Joyful	Insolent	Bored
Capable	Cruel	Feeble
Lucky	Irritated	Nauseated
Responsible	Defiant	Inhibited
Friendly	Hate	Fatigued
Masterful	Hostile	Rejected
Free	Angry	Subdued
Devoted	Annoyed	Impotent
Domineering	Enraged	Ennui
Aroused	Contempt	Blase
Concentrating	Selfish	Haughty
Happy	Reprehensible	Listless
Egotistical	Contemptuous	Deactivated
Carefree	Scornful	Weary
Affectionate	Suspicious	Snobbish
Vigorous	Skeptical	Uninterested
Activated	Burdened	Detached
Alert	Anger	Discontented
Responsibility	Hostile	Discouraged
Controlling	Crushed	Sad
Proud	Frustrated	
Enjoyment	Distressed	
Serious	Insecure	
Cooperative	Humiliated	
Thankful	Hungry	
Respectful	Fearful	
Appreciative	Terrified	
Loved	Embattled	
Grateful	Helpless	
Love	Troubled	
Anxious	Startled	
Impressed	Anguished	
Surprised	Shamed	
Excited	Displeased	
Wonder	Embarassed	
Fascinated	Upset	
Awed	Defeated	

Surrey list of verbal labels:

Love
Joy
Pride
Admiration
Hope
Happy
Relief
Satisfaction
Gratitude
Gratification
Gloat
Pity
Remorse
Shame
Self-reproach
Reproach
Resentment
Disappointment
Distress
Anger
Hate
Fear

Appendix B: Questionnaire

DUBLIN CITY UNIVERSITY

Informed Consent Form

1. **Research Study Title:** Analyzing results of automatic video annotation.
2. **Purpose of the Research:** The study to be carried out is to have human evaluation and feedback on automatically-derived labels from video data.
3. **Procedures:** The study will be carried out in two sessions. The first session will last for approximately one hour and the second session for approximately two hours. The participant will be accompanied by the researcher at all times and is free to take breaks in between. The participant will be required to watch some video clips and answer some questions by filling in a questionnaire while watching the video.
4. **Confirmation that the involvement in the Research Study is voluntary:** Participant is free to choose whether or not to participate in this study and may withdraw from the study at any point.
5. **Privacy:** Participant's identity in this study will be treated as confidential. The results of this study may be published but will not give names or include identifiable references.
6. **Signature:** I have read and understood the information in this form. My questions and concerns have been answered by the researcher. Therefore, I consent to take part in this research study.

Participant's Signature: _____

Name in Block Capitals: _____

Witness: _____

Date: _____

List of films:

1. Alien
2. Chariots of fire
3. Chocolat
4. Clockers
5. Crouching tiger hidden dragon
6. Disco pigs
7. ET
8. Finding Nemo
9. Happiness
10. Harry Potter and the philosophers stone
11. Harry Potter and the prisoner of azkaban
12. Harry Potter and the chamber of secrets
13. Indiana Jones and the last crusade
14. Indiana Jones and the raiders of the lost ark
15. Indiana Jones and the temple of doom
16. Jackie Brown
17. Leon
18. Mean Streets
19. Minority Report
20. Nora
21. Pirates of the Caribbean
22. Pulp Fiction
23. Roger and Me
24. Shrek
25. Star Wars 4
26. Star Wars 5 the empire strikes back
27. Star Wars 6
28. Taxi Driver
29. The Changeling
30. The English Patient
31. The Godfather Part 1
32. Lord of the Rings the fellowship of the ring
33. Lord of the Rings the two towers
34. The Shining
35. The Sons Room
36. Titanic
37. Toy Story
38. Veronica Guerin
39. Who framed roger rabbit

List of film clips:

1. _____
2. _____
3. _____
4. _____
5. _____

Questionnaire

Section 1: Personal details

1.1. Your full name: _____

1.2. Your age: _____

1.3. Your gender, please circle one: Male / Female

1.4. How often do you go to the cinema? Please circle one:

Seldom / 1-2 times a year / Once a month / 2-3 times a month / Once a week / More

1.5. What genre of films do you like? Please rate from 1 (most preferred) to 5 (less preferred):

		Example:
• Action	_____	1
• Drama	_____	2
• Horror/Thriller	_____	3
• Romantic	_____	4
• Comedy	_____	5

Section 2: Movie details

2.1. Choose five movies you're most familiar with from the list of movies available and answer the questions below.

- Movie 1: How familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Movie 2: How familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Movie 3: How familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Movie 4: How familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Movie 5: How familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

2.2. A list of movie clips will be handed to you. Have you seen these films before? Please circle yes or no for each clip:

- Clip 1: Yes / No

If yes then how familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Clip 2: Yes / No

If yes then how familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Clip 3: Yes / No

If yes then how familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Clip 4: Yes / No

If yes then how familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

- Clip 5: Yes / No

If yes then how familiar with the movie are you? Please circle one:

Remember most scenes / Remember some scenes / Remember movie plot only

Section 3: Feedback

3.1. Please rate how much you agree with the valence and arousal curves by quickly looking and browsing a few sections of the movie on a scale of 0-10. Allow yourself up to 5 minutes for each movie.

(0-3 = disagree, 4-6 = slightly agree, 7-10 = strongly agree)

Valence is the range of feeling from positive (pleasure) to negative (displeasure)

Arousal is the range of levels of activity from positive (excited) to negative (sleepy)

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5
Valence curve:	_____	_____	_____	_____	_____
Arousal curve:	_____	_____	_____	_____	_____

3.2. Please watch (or skim through) the video clips. For each clip, write down some words you might use to describe the dominant emotions contained in the video clip. Allow yourself up to 5 minutes for each clip.

Example: Clip 1: intense, full of tension, fear, explosions, shocking, crazy

Clip2: joyful, triumphant, incredible,

- Clip 1: _____
- Clip 2: _____
- Clip 3: _____
- Clip 4: _____
- Clip 5: _____

3.3. You will be presented with a list of automatically generated words to rate on a scale of 0-10 how related the generated list of words are to the original clip. Allow yourself up to 5 minutes for this task.

(0-3 = unrelated, 4-6 = slightly related, 7-10 = strongly related)

Scores:	List 1	List 2	List 3
Clip 1:	_____	_____	_____
Clip 2:	_____	_____	_____
Clip 3:	_____	_____	_____
Clip4:	_____	_____	_____
Clip5:	_____	_____	_____

3.4. That's it for today! See you on the next session.

3.5. (Second session) For each clip, watch (or skim through) the top 10 result clips that were generated, and rate how close each of the result clip is related with the original clip in terms of emotions on a scale of 0-10. Review the emotional labels of the result videos and rate how well these 3 lists of labels relate to the original clip. Allow yourself up to 20 minutes for each clip.

(0-3 = unrelated, 4-6 = slightly related, 7-10 = strongly related)

Example:

		Video score	List 1 Labels score	List 2 Labels score	List 3 Labels score	Comments
Clip1	Result Video 1					
	Result Video 2					
	Result Video 3					
	Result Video 4					
	Result Video 5					
	Result Video 6					
	Result Video 7					
	Result Video 8					
	Result Video 9					
	Result Video 10					

		Video score	List 1 Labels score	List 2 Labels score	List 3 Labels score	Comments
Clip2	Result Video 1					
	Result Video 2					
	Result Video 3					
	Result Video 4					
	Result Video 5					
	Result Video 6					
	Result Video 7					
	Result Video 8					
	Result Video 9					
	Result Video 10					

		Video score	List 1 Labels score	List 2 Labels score	List 3 Labels score	Comments
Clip3	Result Video 1					
	Result Video 2					
	Result Video 3					
	Result Video 4					
	Result Video 5					
	Result Video 6					
	Result Video 7					
	Result Video 8					
	Result Video 9					
	Result Video 10					

		Video score	List 1 Labels score	List 2 Labels score	List 3 Labels score	Comments
Clip4	Result Video 1					
	Result Video 2					
	Result Video 3					
	Result Video 4					
	Result Video 5					
	Result Video 6					
	Result Video 7					
	Result Video 8					
	Result Video 9					
	Result Video 10					

		Video score	List 1 Labels score	List 2 Labels score	List 3 Labels score	Comments
Clip5	Result Video 1					
	Result Video 2					
	Result Video 3					
	Result Video 4					
	Result Video 5					
	Result Video 6					
	Result Video 7					
	Result Video 8					
	Result Video 9					
	Result Video 10					

3.6. Would you like to see the methods of presenting emotions in this study being used as a tool for searching videos in a video search engine one day? Why?

_____

_____

3.7. Please add in any comments you might have:

_____

_____

_____

_____

3.8. This study is finally finished! Thank you for your precious time!

Publications Arising From This Research

Related Publications:

[Chan and Jones, 2005] Chan, C.H., Jones, G.J.F: Annotation of Multimedia Audio Data with Affective Labels for Information Management, *5th International Workshop on Pattern Recognition in Information Systems*, Miami, USA, (PRIS 2005), pp94-103.

[Chan and Jones, 2005] Chan, C.H., Jones, G.J.F.: Affect-Based Indexing and Retrieval of Films, *13th ACM International Conference on Multimedia*, Singapore, (2005).

[Rothwell et al., 2006] Rothwell, S., Lehane, B., Chan, C.H., Smeaton, A.F., O'Connor, N., Jones, G.J.F., Diamond, D.: The CDVPlex Biometric Cinema: Sensing Physiological Responses to Emotional Stimuli in Film, *Proceedings of the 4th International Conference on Pervasive Computing, Dublin, Ireland*, (Pervasive 2006).

Invited Presentation:

[Jones and Chan, 2004] Jones, G.J.F., Chan, C.H.: *Affect-Based Indexing and Retrieval of Films*, Seminar presentation at University of Surrey, Department of Computing, University of Surrey, UK, (2005).