Cinematographic Shot Classification Frameworks for Movie Indexing and Retrieval

A DISSERTATION PRESENTED

BY

MUHAMMAD ABUL HASAN

TO

THE FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

COMPUTER ENGINEERING

UNIVERSITY OF TECHNOLOGY, SYDNEY (UTS)
SYDNEY, NSW

NOVEMBER 2014

Certificate of Original Authorship

Title:	Cinematographic Shot Classification Frameworks for Movie Indexing and Retrieval
Author:	Muhammad Abul Hasan
Degree:	Doctor of Philosophy

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Author

Date: December 5, 2014

Muhammad Abul Hasan
Supervisor: Professor Xiangjian He, Ph.D.
Co-supervisor: Min Xu, Ph.D.

Cinematographic Shot Classification Frameworks for Movie Indexing and Retrieval

Abstract

Cinematographic shot classification is an important and challenging task due to its creation mechanisms. A variety of shot types are used in movies in order to attract audience attention and enhance their viewing experiences. In order to index the cinematographic shots in video databases, shot classification is considered as a primary task. In order to classify cinematographic shots, we propose three frameworks in this thesis. Firstly, we propose a context saliency based framework. In the proposed framework, we introduce context saliency based feature extraction technique from a keyframe of a cinematographic video shot. The extracted features from a training dataset are used to train a Support Vector Machine (SVM) to classify the cinematographic shots into pre-defined shot classes. In the second framework, we propose another keyframe based shot classification technique. In this technique, in addition to context saliency map features, a set of cinematographic domain feature extraction mechanisms are proposed for cinematographic shots classification. The proposed approach works in a hierarchical manner. There are two steps involve in the proposed method. Firstly, shots are classified based on depth information extracted from keyframes. Secondly, shots are further classified by using orientations of objects on keyframes. For classification we use SVM. In the third framework, we propose a non-parametric camera motion descriptor called CAMHID for video shot classification. In the proposed

method, a motion vector field (MVF) is constructed through the extraction of motion vectors using block matching on a sequence of consecutive video frames. Then, each frame is divided into a number of local regions of equal size. Next, the inconsistent/noisy motion vectors in each local region are eliminated through a motion consistency analysis. The remaining motion vectors of each local region in the sequence of consecutive frames are further collected for a compact representation. A matrix is formed using the motion vectors. The matrix is then decomposed using the singular value decomposition (SVD) technique to identify the dominant motion. The angle of the most dominant principal component is then computed and quantised to represent the motion of the local region using a histogram. In order to represent the global camera motion, the local histograms are combined. The effectiveness of the proposed motion descriptor for video shot classification is tested by using SVM. The proposed camera motion descriptor for video shots classification is evaluated on two video datasets consisting of regular camera motion patterns (e.g., pan, zoom, tilt, static). As an application of CAMHID, we extend the camera motion descriptor by adding a set of features for classification of cinematographic shots. The experimental results show that the proposed shot level camera motion descriptor has a strong discriminative capability to classify different camera motion patterns of different videos effectively. We also show that our approach outperforms state-of-the-art methods. Additionally, we further apply CAMHID features in video copy detection task as another application.

Publications from this Thesis

Journal Publications:

[1] Muhammad Abul Hasan, Min Xu, Xiangjian He, and Changsheng Xu. "CAMHID: Camera Motion Histogram Descriptor and Its Application to Cinematographic Shot Classification." *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24 , issue 10, pp. 1682 - 1695, October 2014.

[2] Muhammad Abul Hasan, Min Xu, Xiangjian He and Yi Wang. "A Camera Motion Histogram Descriptor for Video Shot Classification." *Multimedia tools and applications*, DOI. 10.1007/s11042-014-2218-5, 2014.

Conference Publications:

[1] Muhammad Abul Hasan, Min Xu, Xiangjian He, and Ling Chen. "Shot classification using domain specific features for movie management." In *Database Systems for Advanced Applications*, pp. 314-318. Springer Berlin Heidelberg, 2012.

[2] Min Xu, Jinqiao Wang, Muhammad Abul Hasan, Xiangjian He, Changsheng Xu, Hanqing Lu, and Jesse S. Jin. "Using context saliency for movie shot classification." In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 3653-3656. IEEE, 2011.

Contents

Listing of figures

List of Tables

Dedicated to

My Parents

Acknowledgements

I AM GREATLY INDEBTED TO MY SUPERVISOR, Professor Xiangjian He for his continuous encouragement, advice, help and invaluable suggestions. He is such a nice, generous, helpful and kind hearted person. I feel really happy, comfortable and unconstrained with him during my PhD study. I owe my research achievements to his experienced supervision. Many thanks are also due to my co-supervisor, Dr. Min Xu for her valued suggestions and constant support, and for the numerous conversations with her. I gratefully acknowledge the useful discussions with Dr. Ruo Du, Ehsan Zare Borzeshi and Sheng Wang. I wish to thank my fellow research colleagues and the staff of the school, especially those people listed below for providing various assistance for the completion of this research work.

- Professor Massimo Piccardi, Professor Doan B. Hoang, Associate Professor Mao Lin Huang, Associate Professor Valerie Gay, Dr. Qiang Wu, Dr. Wenjing Jia, Dr. Yida Xu, Max Hendricks, Dr. Ruo Du, Dr. Chao Zeng, Sheng Wang, Dr. Zhiyuan Tan, Man To Wang, Mohammed Ambu Saidi, Ehsan Zare Borzeshi, Ava Bargi, Minqi Li, Guopeng Zhang and Wenbo Wang.

I would like to thank my wife, Jannat Sultana, for her understanding and assistance. I also thank my father Mr. Md. Tafazzal Hossain and my mother Mrs. Mahmuda Begum for the freedom to study for the long time necessary to complete postgraduate studies. This thesis could not have been completed without

the support and encouragement of my siblings Mansura Akther, Forhad Hossain, Shohorab Hossain and Umme Salma Aakhi. My special thanks go to my friends Tanveer Ahsan, Tanvir Rahman, Anisul Karim, Jainul Abedin, Abdullah Al Wadud, and Shamsul Alam. Last but not least, the financial assistance of the University of Technology, Sydney International Research Scholarship and Faculty of Engineering and Information Technology Scholarship for my living are gratefully appreciated.

I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right.

Albert Einstein

1 Introduction

1.1 Introduction

Video is an important medium and it describes the visual content of a scene with the help of time domain to human eyes. Smooth visual information flow is achieved by capturing a significantly large number of sequential frames per second to comply with the human brain's cognition speed limit. As a result, video cameras have to capture highly redundant visual data. With the ever increasing production of video data, the demand of efficient indexing, retrieving and browsing is also increasing. Developing such techniques are being considered as one of the major goals in the multimedia data research community [5]. A video produced by a freely moving camera is a rich form of data which is heavily existed in today's multimedia

contents. Apart from the visual information, interesting spatio-temporal information is also buried inside the raw video data. Finding and using such information linking with the semantics is the key to index video data.

Out of wide variety of video contents, movie is one of the most influential media of entertainment for the consumers of all around the world. As the world becomes highly connected where ever-increasing amount of video information is just a click away, it is now a big concern to the multimedia researchers to develop effective techniques so that the videos can be stored and managed efficiently. The recent advancements in video compression technologies and high speed Internet connectivities have made it possible for an easy distribution of videos to the end users. Based on such technologies, many web based video related user applications have been emerging such as online movie databases (i.e., IMDb and Netflix), social video sharing websites (i.e., Youtube and Bing), digital video libraries and video-on-demand. In order to provide effective and efficient access to the video databases, a wide variety of research is taking place. Many video indexing solutions have been proposed for most of the video genres. However, movie domain video indexing and retrieval task is relatively ignored. Due to the redundant form of video data and the internal structure of the movie, developing a movie indexing and retrieval is a challenging task. The challenge mainly comes from the movie making procedure itself. A movie is a collective brainchild of key personnel involved in the making procedure. The key personnel for making a movie include the movie's script writer, director, cinematographers and actors. The artistic involvement of each personnel enables a director to make a movie. A movie

is produced mainly in three major steps: pre-production, production and post-production [6]. The *pre-production* step mainly includes screenplay writing and planning. In the *production* phase, cinematographers make cinematographic shots. In the *post-production* step, directors put the selected shots sequentially along with dialogues and music in an artistic storytelling manner to craft a finished production. At this stage, the directors apply intuitive creativity to create movie for the target audience. This phase is a highly complex stage as the movie post-production is mainly a subjective matter. As a whole, the final outcome of the complete procedure is a sequence of cinematographic shots. Finding the semantics from such an unorganised collections of cinematographic shots and indexing them are a big challenge.

In order to index cinematography shots in a movie database, it is important to identify the points of interests of the movie users. There is a relationship between a point of interest and semantics. Points of interests vary according to users. In any case, shot level movie analysis is considered as a rudimentary task. The structure of a movie can be discovered by alalysing its shots. Moreover, the semantics of each shot can also be discovered by analysing the shot. In order to do that, identifying the characteristics of each shot type and extracting them are the first challenge. Consequently, indexing a movie according to its identified characteristics is next challenge. Traditionally, such characteristic identification is done by using spatial, temporal and/or both features.

In order to dig into the detail of cinematographic shot level analysis, firstly, we identify video shot types from different perspectives. Based on different perspec-

tives, we classify different video shot types. According to the generic characteristics of a video shot type and cinematographic intrinsic characteristics, cinematographic shot types are classified from different points of view. Finally, we discuss about different approaches to cinematographic shot classification.

1.2 Shot Classification

In professional video photography, cameramen often capture video shots from different angles and different depths. Each shot is captured intentionally, and in a grammatically correct way and a proper manner to tell a story. Therefore, each video shot preserves semantics from a director's point of view. Alternatively, it is also true that different types of shots are directly related to a viewer's emotions. Using different shots, directors create affective scenes for movies. A scene is considered as a story unit which consists of a number of consecutive shots. A shot is defined as a continuous strip of visual data which is made up of a series of frames. Figure 1.2.1 shows the syntax of a movie structure. A video shot can be classified based on two basic characteristics. It is based on the distance of an objects from the camera or based on the camera operation. In the following, these two characteristics are discussed in detail.

1.2.1 Shot Type Based on Camera Distance

Depending on the content and emotional involvement of a shot, a camera distance varies significantly. Therefore, the size of an objects of interest on a view plane

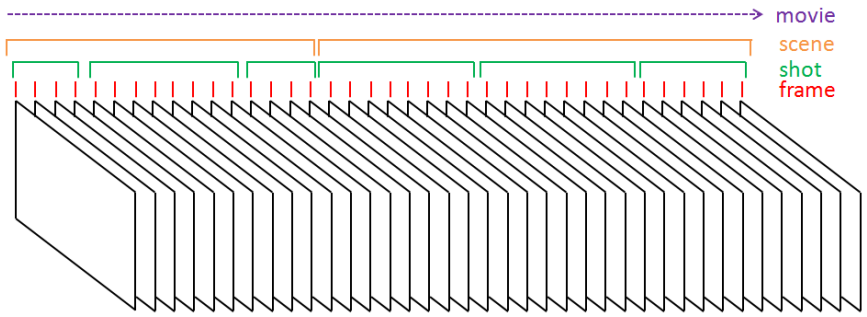


Figure 1.2.1: Syntax of a hierarchical movie structure.

shows a level of emotional involvement. For example, a close-up shot in a movie often shows the facial expressions of actors. A long distance shot is usually used to show the wide view of a location. In terms of emotional involvement, the depth of a shot is related to semantics of the shot. Based on the distance of a focused object from a camera, video shots can primarily be classified into three broad classes: close-up shot, medium shot and long shot [7].

- Close-up shot: This type of shots is one of the standard shots and is frequently used in movies. In this type of shots, the most detailed information of a face or an object is shown by tightly framing in it. Close-up shots can vary greatly ranging from a extremely close view of an object or (e.g. a human eye) to the view of an actor up to his waist level [8]. Figure 1.2.2(a) shows some examples of close-up shots.

- Medium shot: This type of shots is another frequently used in cinematography. It shows a bigger picture of the performers, and often used to relate an actor with context. In this type of shots, an actor's body above his knees is taken. It is often used to show the body language of an actor. Figure 1.2.2(b)

(a) close-up shot

(b) medium shot

(c) long shot

Figure 1.2.2: Examples of video shot types based on camera distance are shown in (a) - (c).

shows some examples of medium shot.

- Long shot: This type of shots shows the whole body of an actor. It provides a big picture of the background surrounding the actor. It is used to establish the locations of shooting-sets. Figure 1.2.2(c) shows some examples of long shot.

Movie directors choose an appropriate depth in each shot for the need of a story.

1.2.2 Shot Type Based on Camera Operation

While shooting, cameramen often use their skills to handle cameras effectively by taking into account the video domain knowledge. Thus, apart from camera distance from performers, shots are to be captured through skilled camera operations.

A set of well defined camera operations are routinely performed to accommodate different actions of the objects of interest on a view plane. According to the camera operations, video shots can be classified as follows [7].

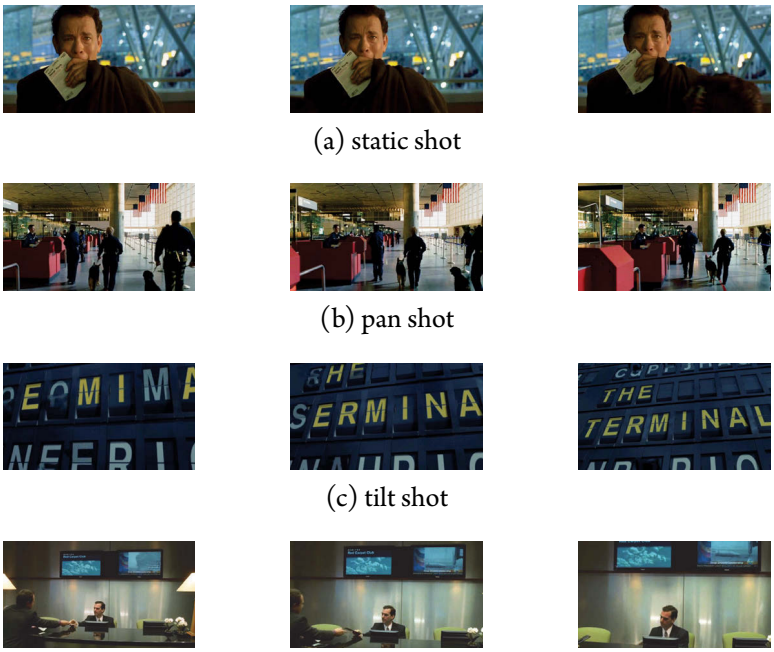
- Static shot: Static shots are captured by placing or holding the camera firmly without any significant camera movements. For close-up view of objects, static video shots are often captured. Figure 1.2.3(a) shows an example of static shot.

- Pan shot: The panning shots are captured by rotating the camera about the vertical axis. For the purpose of following objects horizontally, pan shots are often used. Figure 1.2.3(b) shows an example of pan shot.

- Tilt shot: Tilt shots are captured by rotating the camera about the horizontal axis. Tilt shots are used for following objects in the horizontal direction. Figure 1.2.3(c) shows an example of tilt shot.

- Zoom Shot: Zoom shots are captured by changing the focal length of camera lenses. Depending on the situation, we observe two types of zoom shots: zoom in and zoom out. Figure 1.2.3(d) shows an example of zoom shot.

The above mentioned camera operations are applied according to the need. For example, static shots are often used to display the emotions of actors. Zoom shots are used to increase or decrease tension. Pan or tilt shots are used to follow objects in a scene. In practice, multiple shooting technique may be applied in a single shot.



(a) static shot

(b) pan shot

(c) tilt shot

(d) zoom shot

Figure 1.2.3: Video shot examples based on camera operations are shown in (a) - (d). The left column represents the first frame of the shots. Second column shows an intermediate frame of each shot. The third column shows the last frame of each shot.

However, the shot can be classified based on the majority of the portion's camera motion characteristic.

1.3 Cinematographic shot types and semantic indexing

Film making is completely based on film making grammars. Directors heavily apply these film making grammars on every single cinematographic shot. The directors' main intention is to visualise a screenplay by capturing a cinematographic shot through a set of camera motions and a set of viewpoints. Capturing gram-

matically correct cinematographic shots ensures the viewer's attention on the predetermined actor(s), object(s) or place(s) based on the screenplay. In the following, cinematographic shots are classified based on camera and object positions and based on directing semantic classes.

1.3.1 Cinematographic Shot Types Based on Camera and Object Positions

Apart from the distance from a camera to a performer, movie shots can be further classified based on the actor's position, actor's orientation, size of an actor and content of a shot. Therefore, according to movie theory [8], close distance shots can be further classified into the following sub-classes:

1. Close-up shot: This type of a shot shows a close view of the upper-part of a performer's body. A close-up shot frames a person from his chest to a close view of his face to show his facial expression. Affective contents of movies are mostly represented using close-up shots [8].

2. Over-the-shoulder shot: In a face-to-face conversation scene, a camera is put at an angle from where the back side of the shoulder of an actor is shown in the frame and the camera points towards the other actor's face. This type of shots is a type of the most used shots in movies.

3. Medium close-up shot: Medium close-up shots have a bigger view than other close shots and show performer's body parts above the waist level.

4. Medium shot: Medium shots have a larger view than medium close shots and show performer's body parts above knee level. These shots are good in showing the body language of the actor.

5. Cut shot: Cut shot shows different parts of a human body other than his face. Cut shots are mostly used to bring tension to audience.

There are different types of long distance shots. For example, extreme long shot, medium long shot and long shot. Since the purpose of all long distance shots is to show the location in a scene, we decide not to further classify the long shot into sub-classes. Therefore, we propose to classify movie shots into six classes which are close-up shot, over-the-shoulder shot, medium close-up shot, medium shot, cut shot and long distance shot. The various types of shot classes are used to represent different types of actions/settings which are independent of each other.

Using the dynamic spatio-temporal information, many methods have been proposed for video concept detection and content based video indexing and retrieval [9–13]. In a similar way, a robust semantic movie shot classification technique can be used for many applications such as shot indexing and retrieval [14], movie analysis for understanding the semantics of a shot [15], automatic movie editing for theme representation [16], constructing movie structure for browsing a movie [17], video summarisation [18] and movie genre classification [19]. Although movie shot classification has many potential applications, it is considered as a challenging task. There are many reasons involved, as movie making is considered as a complex and multidisciplinary medium. The approaches based on motion char-

(a) Close-up shot

(b) Medium close-up shot

(c) Over-the-shoulder shot

(d) Medium shot

(e) Cut shot

(f) Long shot

Figure 1.3.1: Example of commonly appeared shots in movie (a) - (f).

acterisation have played an important role in this regard [20].

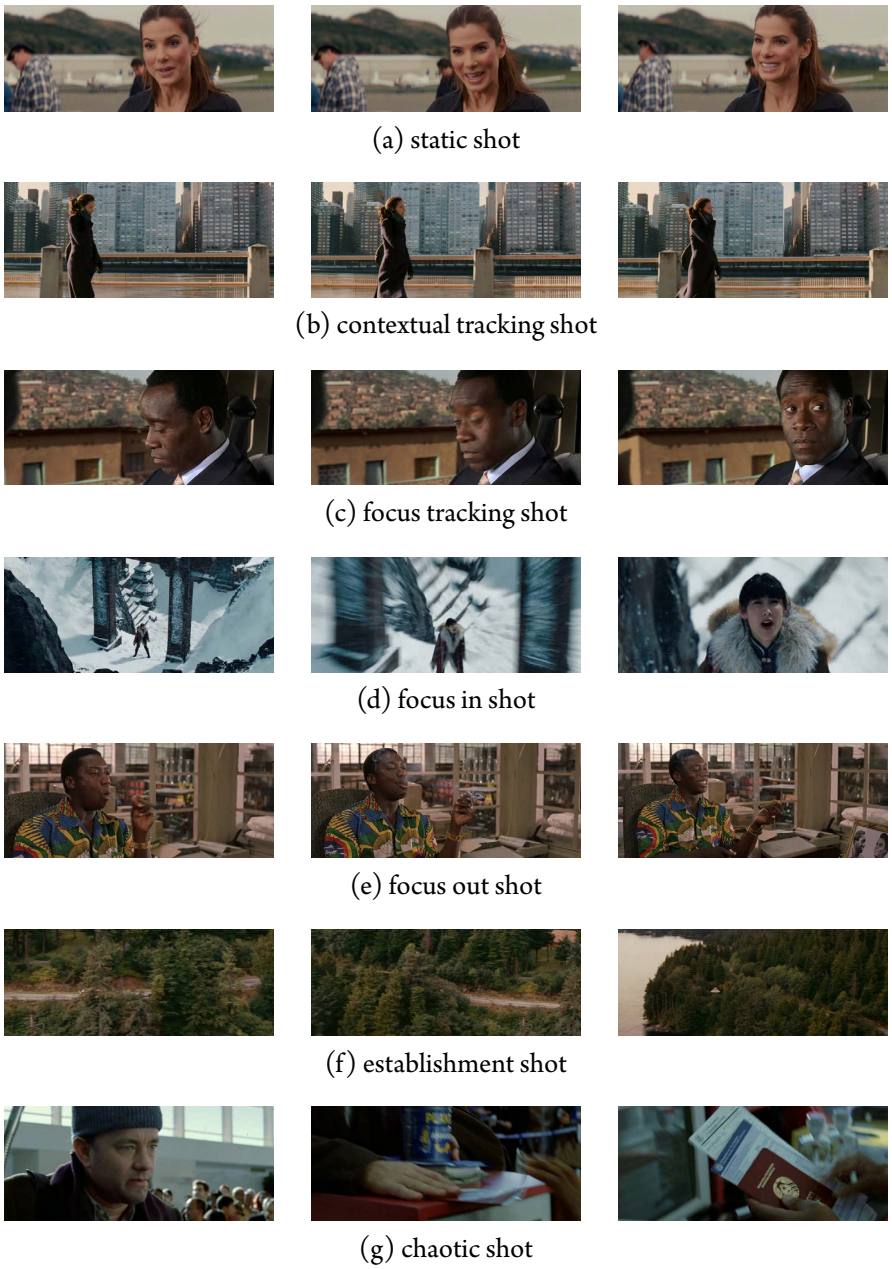
1.3.2 Cinematographic Directing Semantic Classes

Based on camera distance and the camera motion characteristics of cinematographic shots, the construction of a taxonomy of the cinematographic directing semantics is discussed in the following. The relationship of the camera motion and object distance is important in directing semantic classes. The presence of a focused object makes the viewers feel like they are tracking the object. For example, a panning shot with a focused object gives a feeling to the viewers that the viewers personally track the object. However, without any focused object, a panning shot simply shows a place to the viewer. In cinematography, this type of shots is only used to establish a new setting influencing viewers' mind. Scene composition is another aspect of cinematography which handles different issues such as distance of camera, camera angle and light. Among them, distance of camera is crucial as it de-

termines the degree of emotional involvement of a viewer. In movies, we often see that highly emotional scenes are presented by using close-up/medium shots. Long distance shots are used to establish the context of a focused object. For the task of cinematographic shot classification, we group close-up and medium shots into one class as it is not easy to distinguish the purposes of using these two types of shots. In a wide range of movies, the use of close-up and medium shots are very similar for similar emotional shots. However, long shots are mainly used for contextual tracking and contextual establishments. In the following two subsections, we introduce two methods for grouping cinematographic shot types. Firstly, cinematographic shots are classified using the camera and object positions. Secondly, cinematographic shots are classified using cinematographic directing semantics.

The cinematographic directing semantic classes are created using the basic directing elements discussed in the previous subsection. In [4], the cinematographic shots are analysed based on the directing elements and finally grouped into seven semantic classes. 1) stationary, 2) contextual-tracking, 3) focus-tracking, 4) focus-in, 5) focus-out, 6) establishment, and 7) chaotic shots. In reality, the directing semantic classes of cinematographic shots are not clearly categorised into other video domains (e.g., attack shots in soccer video). However, meaningful indexing is still possible using the introduced directing semantic classes. In the following, seven semantic classes are briefly described.

Stationary shot: A significant portion of cinematographic shots are dialogue shots and the dialogue shots are mainly captured by using stationary shots. Stationary shots contain a minimum amount of camera movements to concentrate



(a) static shot

(b) contextual tracking shot

(c) focus tracking shot

(d) focus in shot

(e) focus out shot

(f) establishment shot

(g) chaotic shot

Figure 1.3.2: Cinematographic directing semantic shot classes are shown in (a) - (g). The left column represents the first frames of the shots. Second column shows an intermediate frame of each shot. The third column represents the last frame of each shot.

the viewer attentions on the actor's activities. Figure 1.3.2(a) shows an example of static shots. In this particular case, as it can be seen, the shot is captured by focusing on the actress using close-up shot while the camera movement remains almost static.

Tracking shot: Tracking shots are the one which are captured by focusing on object(s) and follow along the direction of a movement. This type of shots is used to closely relate viewers to the objects [21]. It makes viewers feel like they are following the objects. Because of its own characteristics, tracking shots are considered as an important shot class. There are two types of tracking shots used in cinematography. 1) Contextual tracking shots which establish a relationship of an object with the context by capturing a bigger picture of the scene. The focused object is captured using a long shot so that the object looks smaller but provides scenic detail of the shooting set by using panning camera movements. Figure 1.3.2(b) shows an example of contextual tracking shot where the actress is being shot with a clear indication of the context (cityscape view). 2) Focus tracking is another variant of tracking shots which provides a closer view of the objects. The intention behind taking this type of shots is to focus on the closer detail of the object while tracking. Figure 1.3.2(c) shows an example of focus tracking shots.

Focus-in shot: In cinematography, focus-in shots are captured in two ways: 1) zooming in by shortening the focal length of the camera lens and 2) moving the camera to the object to shorten the camera distance for a closer view of the object. Both of these are mainly used to provide a greater detail of a focused object to highlight some important detail. Figure 1.3.2(d) shows an example of focus-in

shot, where the object is getting bigger by changing the focal length.

Focus-out shot: Focus-out shots are used to detach emotional involvement of viewers from an object or relax the viewers by changing the viewing space. This effect is usually achieved through zooming out or dolly out shots, as the camera gradually moves away from the subject and creates emotional distance. Figure 1.3.2(e) shows an example of focus-out shot by changing the position of camera distance.

Establishment shot: Establishment shots form another important directing semantic class which is used in cinematography. This type of shots is used to introduce a location to establish a relationship with the subsequent shots. This type of shots is often taken by panning the camera without focusing on any particular object. Figure 1.3.2(f) shows an example of establishment shots.

Chaotic shot: This type of shots is characterised by the chaotic movement of the camera to follow an object or an object action. Chaotic shots are the ones which cannot be characterised as anyone of the above mentioned classes. Generally, a random camera motion happens due to focusing on an object's random motion. In order to represent fast action (or motions), directors apply this technique. In this shot type, it is not usually for the fast moving object to dominate viewer attention. Such shots are usually used to represent thrills and used more often in action films. Figure 1.3.2(g) shows an example of chaotic shots.

In this thesis, we propose three frameworks for cinematographic shot classification. The first two frameworks are based on spatial features and these features are extracted from a keyframe taken from a cinematographic shot. Then, we propose a framework which include both spatial and temporal domain features. Finally,

using the third framework, we show the effectiveness in identifying video copies from video databases.

1.4 Thesis Contributions

The contributions of this thesis are summarised as follows.

- We investigate the performance of contextual saliency map on cinematographic shot classification.
- We develop a set of movie domain specific features for improving the shot classification task which is achieved using contextual saliency maps.
- We develop a motion characterisation technique to encode the camera motion in the temporal direction.
- We introduce a shot level descriptor by incorporating the camera motion and the depth of scene information.
- The performance of shot descriptor is investigated to measure the performance in cinematographic environment.
- We also use the motion description and characterisation technique to identify video copy as an application.

1.5 Thesis Organisation

This thesis is organised as follows.

In Chapter 2, we make a detailed literature review on shot classification. Chapter 3 introduces the framework of cinematographic shot classification using film context saliency maps. Chapter 4 describes cinematographic shot classification performance using movie domain specific features. Chapter 5 talks about camera motion based shot classification. Chapter 6 introduces the framework of cinematographic shot classification based on directing semantic classes. Chapter 7 talks about an original method to detect video copy from a video database. Finally, in Chapter 8, we conclude the thesis and address the future works.

Don't let the fear of losing be greater than the excitement of winning.

Robert Kiyosaki

2 Related Work

2.1 Related Work

In a video production environment, a video document is the outcome of a directing process that integrate multiple modalities to describe the script in the form of multimedia. Directors convert a semantic idea into a video document by using the following modalities:

- Visual modality: It is visual information contained in video frames, such as news room, mise-en-scène, sports ground, etc. Visual information can be naturally or artificially created.
- Auditory modality: It is audible information, which contains dialogues, music, commentary and relevant sounds.

- Textual modality: It is a text form of dialogues, also known as subtitles.

For video shot indexing, features are extracted from these three modalities. In the literature, a diverse variety of solutions have been proposed for wide variety of video genre indexing. Video shots are indexed based on different semantic meanings for different indexing schemes such as, genre-, sub-genre-, motion-, activity- and action based indexing [13]. By now, densely explored solutions have proposed for news, commercial videos and sports videos [13]. In comparison, however, indexing for movies based on different semantics has been less explored. In the following sections, firstly, we discuss about the background and the state-of-the-art shot indexing for different video genres based on different modalities. Then, we discuss about the background and the state-of-the-art techniques for cinematographic video shot indexing.

2.2 Video Indexing Based on Different Features

A video consists of a number of story units and is often termed as scene. A scene consists of a number of video shots taken from different viewpoints and different angles. For simplistic representation, a shot can be represented by one or more keyframes. In the following, we discuss about video shot indexing by extracting different features either from a keyframes or from a shot.

2.2.1 Visual Modality Based Video Indexing

In order to represent a video unit, keyframes are often used. In that case, features are extracted from a representative frame of a video shot. The extracted features

are similar to the content based image indexing and retrieval systems, where traditional image representation techniques are mainly used. In the following, we summarise different visual features extracted from keyframes for video shot indexing problem.

Colour features have been widely used in traditional image retrieval systems. Such features include colour histogram and colour moments on different colour spaces (*e.g.*, RGB, YCbCr, HSI and YUV). Based on the applications, these features are extracted globally (*i.e.*, from the entire keyframe) or locally (*i.e.*, from a particular local region of the keyframe). The advantage of colour features is that it is easy to extract these features with low computational complexity. Additionally, colour features represent human visual perceptions. In [22], colour features were used for concept detection and video retrieval. The colour histogram and colour moment were used in the proposed method. In [23], local colour histograms and colour moments were used as features for video retrieval task. A video search engine was proposed in [24], where colour correlograms were used for video retrieval. In the above works, it was found that only colour based features were not sufficient to represent a keyframe. Along with colour features, some other salient features are equally important in order to represent a keyframe. The colour feature importance depends on the video domains and their applications. In sports video, the type of shots are often simple, at least from the narrative point of view. In soccer video, the difference between shot types is useful for indexing, such as, identifying play time and break time [25], play summarisation [26], and region of interest detection [25]. There are numerous other such applications in the literature for soc-

cer video shot classification. Based on visual modality, many solutions were proposed. In [27], shots were classified by identifying the ratio of the green grass area in a frame. First, the colour of grass field was learned by doing histogram analysis. Then, the dominant hue value was considered as the ground colour to classify the soccer video shots into close-up, medium and wide shots. In order to do that, the pixels belonging to a grass area were identified by measuring the similarity to the learned ground colour. By using Hidden Markove Model (HMM), the dominant ground colour was modelled and two thresholds were determined. The shots containing a large amount (greater than a threshold value) of green colour pixels were considered as wide shots. Shots containing small amount (smaller than a threshold value) of green colour pixels were classified as close-up shots. The rest of the shots were classified as medium shots. Finally, the play times and break times were determined using heuristics for the purpose of indexing and retrieval. Similar approaches were taken in several other sports video shot classification methods [28, 29]. Although, colour learning based shot classification approaches are effective in sports video domain, they suffers from many unanswerable challenges for a generic video shot indexing technique.

Image texture measurements quantify the spatial arrangement of colour or intensities in an image or part of an image. Texture features contain some important information about spatial arrangement of different object surfaces. Texture features include orientation features, wavelet features, Tamura features and co-occurrence matrices. In [22], Tamura features and co-occurrence matrix were used for the TRECVid-2003 video retrieval task. In [30], the Gabor wavelet was

used to capture the texture information for video search engine. The authors used mean and variance of 12 energy oriented filter output to create a texture feature vector. Another Gabor filter based approach was proposed in [31], where local texture features were computed by segmenting the keyframe into a number of equal sized regions. The texture features can be applied effectively where texture information is salient.

Object shapes are also taken into account while extracting features from a keyframe. It describes the shape of an object in a keyframe. In general, the edge of an object is detected and then the detected edge is described by using an edge descriptor. Hauptmann *et al.* used an oriented edge histogram as a shape descriptor in [30]. In [32, 33], local shape descriptors were computed using an oriented edge histogram by segmenting the keyframe in a number of equal sized regions. In general, shape features are difficult to extract. However, they are effective when shape information is salient.

In [34], Gong *et al.* proposed an automatic soccer video parsing technique to classify soccer video shots based on a priori model of line mark detection and recognition in soccer play field. Different marks on the playground, such as center field and goal mouth, were detected using a priori model. This technique is based on game domain specific knowledge and model specific techniques. The proposed technique could classify a sequence of soccer frames into various categories. Although the proposed technique can be applied to other sports video domains, such as basketball and tennis, it is not possible to apply such a technique for general video shot indexing.

The depth of a shot is used as another feature to classify shot types. In [35], the depth of a scene was estimated from the image structure. The absolute depth of the scene was estimated by using local and global spectral features of the image. By using statistical pattern recognition techniques, the absolute average depth was estimated for a given single image. The authors used the frequency domain to discover the global spectral signature of an image. The global spectrum denotes the mean amplitude spectrum which is closely related to the average absolute depth. The expectation maximisation (EM) algorithm was used to find the conditional probability density function (PDF) of the mean depth. Finally, the mean depth was estimated using the PDF according to a mixture model of linear regressions. Another alternative way of determining the depth of a scene is measuring the sizes of the recognisable objects, such as faces, cars and hands contained in a frame [36]. Although depth estimation based approaches are considered as good approaches in identifying depth based shot classes, they are considered as computationally expensive methods for video data processing.

2.2.2 Motion Feature Based Shot Indexing

In the context of motion analysis for video shot indexing and retrieval, there are a great deal of research work accomplished [3, 37–44]. Unlike still image, a video shot integrates motion information to represent the real dynamics in a scene. Motion can happen due to the camera movement or due to object motion. Therefore, in the literature, motion features are divided into two categories: object-motion-based and camera-motion-based features. Object-motion-based features can be

further classified into statistics based, trajectory based and objects' spatial relationship based features. Camera-motion-based features are further classified into static, pan left or right, tilt up or down and zoom in or out features. In the following, object-motion-based indexing and camera-motion-based indexing are discussed.

Object-Motion-Based Indexing

Object motion based features are modelled in different ways in the literature. Statistical based feature modelling is one of the prominent ways to do so. Object motion statistical features are modelled to estimate local and global motion distributions. In [45], a casual Gibbs model was used to estimate spatio-temporal motion distribution. After that, a typical statistical framework was developed for video indexing and retrieval task based on object motion. In [46], motion vector fields were transformed to a number of directional slices to form features called motion textures. The motion textures were then used for video indexing and retrieval task. Although the statistics based motion features are used successfully in some video indexing and retrieval methods, they have limitations to represent the accurate actions of the objects.

Motion trajectories have also been used to model object motion. Using motion trajectories, motions are modelled to create trajectory-based features [47]. In [20, 48], an online video retrieval system was proposed based on object motion trajectories. In [48], motion trajectories were represented as PCA (principal component analysis) coefficients of the temporal orderings of sub-trajectories. In [49], a polynomial curve fitting based motion model was proposed. The mo-

tion model was used as the searching key for object retrieval purpose. In [50], another trajectory based motion model was created by using the motion vectors of the MPEG bit-streams. A similarity measurement based retrieval system was proposed in their work. In [51], trajectories were represented using a series of semantic symbols. Then, a combination of visual distance and edit distance was proposed to measure similarity for video retrieval. Although trajectory based approaches were successfully applied for video indexing and retrieval task, the success is highly dependent on the accuracy of object segmentation. An automatic solution for that purpose is considered as a very challenging task.

Spatial relationships between objects are also used as object based motion features. In [52], a symbolic representation of the spatial relationship between objects was applied for video indexing and retrieval. In [53], each object's spatio-temporal relationship was specified for video indexing and retrieval. Although it is very difficult to identify each moving object in a shot, the advantage of object's relationship based features is that multiple objects' spatio temporal relationships can be represented for indexing video shots.

Camera Motion Based Shot Indexing

In the case of semantic analysis of video shots, camera motion patterns are often used as an important clue. While watching a video, human visual systems can perceive motions which can be described in terms of motion quality: slow or fast motion. However, in a computer vision system, it is described using activity descriptors. In [54], visual motion descriptors were organised into four categories: mo-

tion activity [44, 45], camera motion [55–57], mosaic [58] and motion trajectory [20]. Camera motion descriptors are used to represent the type of camera motion happened in a video shot. Mosaic is captured using the parameters of a parametric motion model of camera. Motion trajectory, as described, tells the object motion in time. Generally, the camera motion descriptors tell the generic inherent camera motion which is used to identify the intention of video shot directors.

Many parametric methods for camera motion detection have been proposed. Using two consecutive video frames of a video, global motion models were proposed in [55, 58–61]. In each case, dominant motion patterns were determined using robust statistical techniques. Although global camera motion detection techniques are theoretically sound, it is considered less feasible to estimate correct parameters in wide variety of videos. The 2D parametric transformation suffers from a weak assumption that the camera distance from the scene is far. This assumption will lead to estimate wrong rotational and translational parameters. In [60, 61], the depth problem was handled through identifying the horizon lines. It is based on the assumption that there is a horizon line present in a typical outdoor scene videos which can be identified using a gradient analysis in gray scale images. This assumption will lead to many wrong parameter calculations where the horizon line is not obvious. None of the mentioned parametric model can be applied in classifying video shots based on directing semantic classes, as directing semantic video shots consist of wide variety of shooting technique in wide variety of shooting sets.

In contrast, nonparametric methods analyse video data using statistical methods to measure local or global camera motions. In [44, 62], the motion distribu-

tion of a shot was represented by using a histogram to analyse the camera motion and video shot similarity measurement. A template matching based approach was used to recognise camera motion in [40, 63]. In [41], a nonparametric spatio-temporal mode-seeking method was proposed in the motion space. The spatial distribution of the dominant motion modes was used to represent motion characteristics. Although this method is capable of learning the semantic concepts of video shots, it does not have the capability to model temporal motion pattern. Ma *et al.* [64] proposed another nonparametric generic motion pattern descriptor for video shot classification. A mapping technique based on a unit circle was applied to transform MV fields to a multidimensional motion pattern descriptor. Although it introduces a temporal information accumulating technique for statistical learning, it lacks an unified representation framework.

Template based camera motion detection was used in [40, 42]. Lan *et al.* proposed a framework for home video camera motion analysis in [42]. A template based background motion estimation (ME) technique was applied to characterise different camera motions. Lee *et al.* proposed an MPEG video stream shot classification technique [40]. In the proposed technique, a video was divided into shots and the shots were classified into six basic camera movements using templates. Template based techniques cannot be used in highly dynamic video sequences due to high random motions captured from different objects.

In order to use motion information buried in the video data effectively, it is important to extract motion information from the frame sequences. In [3, 43], MPEG video MVs were directly used in a camera motion descriptor. Although a

good classification accuracy was achieved, direct use of MVs for compressed videos could be misleading. For the purpose of best representation of a video frame, MVs in a MPEG frame were predicted either from its previous frame (i.e., P-frame) or bidirectionally predicted frame (i.e., B-frame). Bidirectionally predicted frames are much complex in nature as they use both previous and future frames for motion compensation. Thus, it can be concluded that the MVs in MPEG videos do not represent the optimal optical displacement of a macroblock (MB) with respect to time. In [43], backward predicted MVs were mapped to a forward predicted one. Moreover, I-frame's MVs were estimated by interpolation of two nearest P-frames. The overall procedure may produce misleading motion information and eventually may identify a wrong camera motion. In [3], MVs were only estimated from P-frames for characterisation of camera motion. This technique of extracting motion information may suffer from lack of information and hence the estimated camera motion may not be accurate.

2.2.3 Auditory Modality Based Indexing

In addition to motion as a temporal domain feature, audio is also a widely used modality. There are two types of audio-domain features that can be used for video classification.The first type of features is time-domain audio features, which includes RMS value of the audio signal energy, zero crossing rate (ZCR), and silence ratio. Zhang and Kuo [65] proposed an audiovisual content analysis using audio based features. Apart from time domain audio features, frequency-domain features such as the fundamental frequency, the frequency bandwidth and Mel

frequency cepstral coefficients (MFCC) can be used in video classification [66]. In comparison to image/video based features, it is easier to extract audio-based features. However, in [67], it is discussed that video shot classification with only audio-based features is not sufficient.

2.2.4 Textual Modality Based Indexing

Text in the video has been used as another important source of high-level semantics. A text detection based indexing and retrieval system was proposed in [68]. In order to enable and enhance segmentation performance, the proposed method used the typical characteristics of text in videos. First, the superimposed text regions were identified and segmented. Then, using the segmentation output, an OCR system was applied to detect the text for indexing and retrieval. Although the proposed method could be used for some specific types of videos, where text appears more frequently on the cinematographic frames with the sub-titles superimposed. However, in general, such things are not common in feature films or other video genres. Therefore, such technique can not be used for general purpose video indexing and retrieval.

In [69], textual modality was also used besides audio and visual information. A three layer approach was proposed to process low, mid and high level information. Low level features include colour, shape, zero crossing rate and transcript. Mid level features include faces, speech and keywords. High level features include semantic indexing which was computed by using mid level features across different modalities. The proposed method was able to classify a video into talk show,

commercial or financial news.

In [70], a text and face related observation was made in genre dependent videos for classifying videos into news, television commercials, sports, and movies/television series. According to their observations, news videos contain named events, annotation of people and setting related texts; Commercial videos contain product names, claims and disclaimers; Sports videos contain players name and game statistics; and movies contain captions and credits. It is also claimed that the facial appearance also follows patterns in different videos. Based on the face and text related features, each frame is assigned a label from a set of predefined labels. The labels were used as input to a hidden Markov model (HMM) to classify input videos into the above mentioned classes. In [71], a subtitle analysis based video classification technique was proposed. The proposed method assigns a category lable to each video by analysing the appeared words in a subtitle. In [72], a multimodal video summarisation approach was proposed. Among other features, subtitles are included to summarise videos.

2.3 Video Indexing Based on Different Semantics

In the following, we discuss about video indexing techniques based on different semantics, such as logical unit detection, video genre detection and event detection from videos.

2.3.1 Logical Unit Detection

A logical unit in a video is a part of the whole video containing interrelated visual information and as a whole making a complete sense. The news and sports videos are instances of structured video. The logical unit detection in a news video is a straight forward task. The anchor shots can be modelled easily as there is only one person staying in the scene with a minimal movement. In [73, 74], the size of the detected face and a restricted position were used to detect an anchor shot. The visual similarity is exploited in [75–77] to detect anchor shots in news videos. In [75], a motion activity threshold based approach was used to classify anchor shot from the candidates. Based on the motion quantity, a shot is either classified as anchor shot or report shot in a news video. In [77], face and lip movement directions were used as features to classify report and anchor shots. In [76], silence intervals were used as the boundaries of the reports. In [78], a detailed news video indexing technique was proposed. The news videos were labelled with six logical units, namely, begin, end, anchor, interview, report and weather forecast. Each logical unit was modelled by using HMM.

In [79], a soccer video logical unit identification method was proposed. Based on the features, the logical units were classified into play time or break time. A grass colour ratio based detection method was applied to classify a frame into global, zoom-in and close-up shot. Then, play time and break time were identified based on a set of heuristics.

2.3.2 Genre Detection

Video genre detection is another kind of video semantic detection. The shot length statistics are used for indexing in this regard. The length of a shot indicates the pace of the video document. A longer video shot means a slower video document. A news video is considered as a slow video document whereas a commercial TV video is considered as a fast video document. The shot change rate in a commercial TV video is much faster than the news video. In [80], the shot change rate and presence of black frame are used as features to identify the commercial TV videos within the news video. The logic behind identifying the black is that the black frames are used before and after the commercial TV videos for a very short period of time. Identifying such frames indicates the presence of a commercial video. Then, the shot change rate within the boundary was estimated to determine the commercial video. A similar method was used in [81] to identify the commercial videos within feature films. Two more additional features (*i.e.*, edge change ratio and motion vector length) were used along with black frame and shot change rate to represent the higher action in commercial video. In [82], videos were classified into news, commercials videos, cartoons, sports, music by using average shot length, ratio of different transition patterns and six visual features.

Although a generic solution for sports video detection is a very difficult task due to wide variety of sports types, a main stream for sports video detection has been proposed in [83]. Presence of slow-motion replays, presence of some specific camera or object motion, and presence of of overlayed text were used as features. In this work, motion features including motion magnitude help achieve highly accu-

rate results.

Multimodal approaches for video genre classification were reported in [69, 84, 85]. The method, proposed in [84], classified videos into basketball game, football game, news report, commercial video and weather forecast by feeding audio and visual features into a HMM classifier. In [85], a three step approach was proposed for film genre recognition. Initially, colour statistics, motion vectors and audio statistics were computed as content features. Then, video layout features, such as shot length and camera motion, were extracted. Finally, classification was made by using the style profile.

In [86], a method was proposed to identify four different types of sports videos (*i.e.*, basketball, ice hockey, soccer, and volleyball). Motion features were extracted from the consecutive frames and then features were reduced by using principal component analysis. The features were used by two statistical learning methods. Experimentally, it was found that a continuous observation density Markov model produced better results. Although the performance was good, the authors intentionally excluded crowd scenes and time-out scenes in their experiments. A similar method for sports video detection was proposed in [84].

2.3.3 Events Detection

A good amount of event detection works were accomplished by using multimodal features. In [87], a three layer event detection algorithm was proposed for animal hunt event detection in a wildlife video document. The first layer extracts low level features, such as colour, texture and motion features, and detects moving objects.

A neural network was applied in the second layer to determine if a blob is an object of interest. Finally, the target event is detected based on the extracted shot descriptor. In [88], car crashes and violent events were detected by using sound features from feature films. First of all, different chaotic sounds, such as explosions, horns, engine's sound and gunfire, were detected. Then, a video segment was labelled with a high level named event based on the concentration of those features.

In [77], various events were detected from broadcast news videos. The detected events include walking shots, gathering shots and computer graphics shots. Walking shots were detected by using up and down periodic motion of human face. Gathering shots were identified by detecting human faces of similar size. Computer graphics shots were detected by identifying the consecutive frames that lack motion.

In [89], slow motion replays were detected to identify important events in sports videos. The slow motion replays were detected by HMM. Slow motion was used to detect important events in a sports video. In [90], presence of flash lights was detected to identify various events in feature films. Supernatural power, crisis, terror, excitement and general events of importance were detected based on the average luminance of the frame influenced by flash lights.

Change of scores is considered as the most important events in sports videos. In [91], visual and textual modalities were used to find a link in identifying change of scores in American football games. The real-time closed caption was analysed to detect a series of keywords related to an event. At the same time, the visual stream was also analysed and linked with the textual feature to determine the most likely

score event. Similarly, a few articles are found in the literature to detect scoring events in tennis games [92–94], soccer games [34, 95], baseball games [94, 96] and basketball games [97, 98]. Typically, the visual features, including colour, edge and texture, and motion features are used to identify scoring events in games. A different modality is used in [96] to identify the exciting events to generate a highlight of baseball games. In order to do that, sound features are only used. Excited speech of the running commentary and the baseball pitch and hit sound are used in the proposed method.

Duan *et al.* [99] use various semantics categories describing sports video shots for mid-level representation to facilitate high-level analysis. In their work, they presented a generic mid-level representation framework for semantic sports video analysis. The mid-level representation was introduced between the low-level audiovisual processing and high-level semantic analysis. For a robust low-level feature analysis, a non-parametric clustering technique was used in both colour and motion feature analysis. Using this framework, it was possible to detect a big number of sports events having strong semantic meanings. Although it works efficiently and effectively in sports video domain, it is not possible to apply such a technique for video shot classification in other domains. As general video shots do not have any dominant colour clue or well defined structured motion, a mid level representation of colour and motion features is a challenging task.

2.3.4 Various Applications Based on Motion Features

Motion features are important temporal features which were used in a wide variety of applications proposed in the literature. In [100], a camera motion estimation method was introduced for the purpose of annotating a basketball video in MPEG domain. The camera motion was estimated directly from the information obtained in MPEG motion vector fields. By analysing directional panning, a semantic annotation was generated. Since motion vectors for MPEG compression take into account only inter-frame differences, they do not necessarily ensure the obtained displacement is the true representation of an optical flow.

Another work, reported in [101], proposed a framework to analyse the temporal structure of a live sports video. The method was based on colour filtering, object segmentation and edge based verification. This approach depended on visual cues, such as colour, motion, and object layout. Since sports videos are captured in simplified and deterministic places, it is useful to perform frame-level analysis. However, it is not easy to distinguish dynamic characteristics in an individual motion vector field, which could be contaminated.

A shot-level parsing techniques was proposed in [102]. Firstly, tensor histogram features ware extracted from two-dimensional (2-D) temporal images and they encode the motion information. Then, both tensor and colour histograms were used for constructing a two-level hierarchical clustering structure. Each top level cluster contained shots with similar colour. Each bottom level cluster consisted of shots with similar motion. The constructed hierarchical structure was then used for the cluster-based shot retrieval. This method is particularly useful for sports videos of

which motion and colour are important visual cues.

In [45, 103, 104], shot level motion features were used for shot classification. In [105], shots were classified into pan, tilt and zoom using simple shot-level camera motion. Ho *et al.* [106] measured the similarity using motion information within a region of interest. Using this approach, it is difficult to extend such work when there is no particular region of interest. In case of video indexing tasks, motion feature can play a big role. In [107], the statistical properties of transition between camera motion types (pan, tilt, zoom and shake) were used to identify different genres of sports videos (*e.g.*, soccer, baseball and tennis). In [108], various camera motion properties, such as angle of a motion, speed, and number of stages in a camera motion, were derived to detect offensive plays in an American football game. These techniques are dependent only on the domain of sports videos.

A maximum entropy method for an automatic shot understanding and summarisation of a baseball game video was used in [109]. The proposed algorithm creates complete indexes of a baseball game. A multi-modal integration of image, audio and speech features was used to compute a maximum entropy. This technique is also highly dependent on domain knowledge.

2.4 Cinematographic Shot Indexing

Story unit detection in video documents is a widely researched topic. An overview of logical story unit detection method was presented in [110]. For the purpose of indexing, logical story unit detection alone is not enough. The semantic labels of the story units are more important. As discussed in the previous sections, a good

amount of video indexing work have been done. Compared to those, cinematographic video indexing is relatively ignored. In [111], a dialogue scene detection method from the feature film was proposed. In the proposed method, face detection, face location and audio features were used. With the help of HMM, a scene was classified as establishing scene, transitional scene or dialogue scene. In [112], a method was proposed to identify violent scenes in feature films and television dramas. The spatio-temporal activities were measured to determine the level of action contained in a scene. Then, using a predefined colour table, flame and blood were identified from the scenes. The audio channel further provided a clue of the presence of violence. Finally, a knowledge based combination of feature values was used as a classifier. In [113], a multimodal integration method was proposed to identify four types of scenes. Firstly, the sound signals were segmented and labelled with silence, speech, music and miscellaneous sounds. These semantic labels were then combined with a temporal stream of video data. An alternating pattern of detected speech labels indicated a dialogue scene. Repetition of visual patterns indicated a story. A progressive pattern in the visual information and non-speech label in the audio segment indicated an action scene. The rest of the scenes were identified as generic scenes. In contrast, only visual information based approach was proposed in [114], where video scenes were detected as dialogues, actions and story units. Firstly, shots having a similar visual pattern were given an arbitrary label. Then, based on the scene patterns, they were labelled as dialogues, actions or story units.

A few significant works are found related to cinematographic shot classification

in the literature [4, 115–118]. The works in [4, 115] proposed a systematic approach based on motion descriptors to build taxonomy for film directing semantics, where the camera distance from the focus of attention was used as an intermediate feature to distinguish contextual tracking and focus tracking shots. In their work, a Markov Random Field [119] based algorithm was formulated for motion segmentation to extract salient motion descriptors for identifying different types of cinematographic shots based on directing semantics. In the proposed method, a coherent taxonomy of cinematographic directing semantics based on directing tasks were performed by using camera motion behaviour and camera distance estimation. The proposed method could be used for some applications, such as video content management and processing. However, for some applications, such as semantic movie indexing and retrieval, content based analysis is highly desired.

In [120], a Lie algebra based cinematographic shot classification technique was proposed to classify shots into aerial, bird-eye, crane, dolly, establishing, pan, tilt and zoom shots. In the proposed method, a homography was assumed to existed between a pair of subsequent frames. By using image based methods, the homography parameters were estimated to represent course camera motion. Then, the homography matrices were mapped to a vector space. Then, all the vectors were stacked in a vector time-series. Finally, an efficient linear dynamic system was used to extract meaningful features from the vector time series. The features were further used to train SVMs for classifying different shot classes. This method is considered to be powerful for video shot indexing. However, for video shot retrieval, this method cannot be used directly.

In [118], human body based rules were applied to identify shot types in a small set of data. In movies, the sizes of a human faces often determine shot types. For example, if the size of a face is big, the possible shot type is of close shot. Similarly, if the size of a face is very small, the possible shot type is of long shot. Based on these observations, a number of thresholds were used to estimate the dimensions and positions of faces in video frames for classifying shot types. Although there have been some merits used to identify shots using the proposed method, it fails to classify generalised cinematographic shots. For example, this method does not work well in identifying cut shot and over the shoulder shot.

In [117], five different inherent characteristics were used to estimate the camera distance from the main subject. The first descriptor was called local distribution of colour intensity which measured the percentage of background pixel with respect to the frame area. The descriptor was computed using keyframes of cinematographic shots. The algorithm developd a method to estimate camera distances. A second order statistics was performed on the keyframe's histograms. Using the information, histogram variance image was created. Based on the created histogram variance image, background and foreground pixels are identified. The percentage of foreground pixels has a good indication of the camera distance from the subject. If the percentage is high, then the image is declared as a close up image. A feature for camera distance estimation was computed by using a motion descriptor. By using a motion activity map, foreground and background motions were identified. In a similar way, foreground and background regions were identified. The above estimation indicates a distance measure from the camera. A descriptor related to

the camera distance was obtained by measuring the scene perspective. A Haugh transformation based perspective detection technique was described in their work. Face size was used as an important clue of the distance of an object from the camera. A face detector was used to identify faces in the cinematographic shots. The camera distance was estimated based on the sizes of faces. The proposed method only classifies cinematographic shots into only three types of shots (i.e., close-up shot, medium shot and long shot). It does not provide any semantic clue of the intrinsic characteristics of the shots.

Efficient and effective handling of video documents depends on the availability of indexes. It is obvious that manual shot classification or indexing is not feasible for large video collections. However, for an effective indexing task, a feature combining features in both spatial and temporal domains should be used for classification. Therefore, instead of separately treating the different feature sources involved and specific algorithms for feature extraction, we need to focus on the relationship of different types of features. In that sense, in our approaches we use both spatial and temporal features for better representations of cinematographic shot types. In the first two frameworks proposed in this thesis, we use spatial domain features only. In the third framework, motion features are included.

2.5 Discussion

We have discussed about the state-of-the-art of video indexing and retrieval from the literature from a broad point of view. Although we have discussed a wide variety of video genres, we restrict ourselves to work only on the video shots in the

film domain (i.e., cinematographic shots). For doing this, we have reviewed different modalities and their applications for various video indexing purposes. To show a comprehensive view, we pay attention to the most discriminating properties of shots corresponding to different video genres. Among the video genres, talk shows have very limited settings with a well defined structure. News videos are another kind of videos with limited settings. In news videos, anchor shots are designed with limited settings and they have discriminating features able to differentiate different shot types in high accuracy. Music videos are not considered as well defined video documents, and the need of indexing a music video is very limited. The reason is that a music video mainly follows a music track and the whole video is directly related to the music. Sports videos are also considered to be simply structured video documents. However, a wide variety of sports genres have made it difficult to develop a generic technique for sports video indexing. Documentaries are characterised by their slow pace. Moreover, the presence of a narrator's voice helps to distinctly identify this video genre. In contrast, feature films and drama videos have complex layouts and properties. These video documents are mainly showing human dialogues and a wide variety of actions. In order to provide an effective viewing experience, a wide variety of shooting techniques are applied. These result in a complex structure compared to the other types of video documents.

In conclusion, although many research efforts have been made to characterise camera motions of video shots, there is still a room to improve the results. Moreover, cinematographic shots have been relatively ignored. An efficient motion rep-

resentation and an effective motion descriptor can represent a video shot for the purpose of semantic video indexing.

In three words I can sum up everything I've learned about life:
it goes on.

Robert Frost.

3 Saliency Based Shot Classification

3.1 Introduction

In this chapter, we propose an original framework for cinematographic shot classification by identifying the visually salient locations on the keyframes of a cinematographic shots. The high level concept of salient object is incorporated in the context of movies to compute the visually conspicuous locations. Context saliency is an extension of traditional visual saliency identification techniques. In this chapter, our motivation is to investigate the performance of the context saliency features along with colour intensity and texture distribution features in classifying cinematographic shot. The main contributions of this chapter are as follows:

1. In a novel effort, we investigate the performance of the context saliency

map in classifying cinematographic shots. Context saliency map is a refined saliency may which not only consider the saliencies form different objects but also consider the context of the surrounding environments. The proposed context saliency map is used to identify the salient regions in a cinematographic shot.

2. The proposed method is movie genres independent. Thus, it is capable of classifying a movie shot irrespective to any movie genre.

3. The proposed method classifies shots into the most frequently appearing types of cinematographic shots.

Visual attention has been proved meaningful for image and video analysis [121], especially for video highlight detection [18, 122] and image summarisation [123]. It is considered worthy to investigate the performance of context saliency maps in classifying cinematographic shots. We investigate the use of visual attention distributions on a movie screen with colour intensity and texture distribution information to infer cinematographic shot type. In Section 3.2, we review the visual attention models. Section 3.3 describes the context saliency map generation technique. Section 3.4 describes local and global features extraction methods. In Section 3.5, experimental results are presented.

3.2 Review of the Visual Attention Models

Although there are a number of theories on the primates' visual attention systems available in the literature, the mechanism of the visual attention systems of the

primates is yet to be fully understood. The research efforts of finding the updated theories behind such mechanism are progressing continuously. Based on the outcomes of the research on visual attention systems during the past few decades, there are a number of functional frameworks derived. The inspiration of the visual attention models to identify the conspicuous location on an image is originated in cognitive psychology and neuroscience theories of human visual attention. The major motivations for the computational visual attention development are:

- to check the legitimacy of the theories of visual attention described in psychology and neuroscience literatures; and
- to apply the visual attention theories in different applications of computer vision fields.

There are two distinct research trends in computational visual attention modelling [124]. The first trend is investigating the neuronal responses by simulating visual systems of the primates during various attentional activities [125–128]. The second trend is to utilise the unique properties of biological attention systems to develop a technical visual attention system [129–133].

The earliest research effort on visual attention systems was proposed by Treisman *et al.* in their famous feature integration theory [134]. The broad concept of feature integration theory is that the objects are perceived by the primates by perceiving and combining different features of the objects. Firstly, various features of the objects integrate themselves to the visual system. Then, with the help of primates' visual attention, the concept of object or action is recognised. In order to

do that computationally, various visual features, such as colour, brightness, orientation, texture, and spatial frequency are used to compute separate feature-maps. Then, all the feature maps are combined to form the final-map which is known as saliency map. The saliency maps are used to direct the attentions based on the importance of the computed local saliency values.

According to [124], some visual attention models are modelled targeting a specific application. In such models, there are two basic properties used to tune the performance of the desired application. The properties are as follows.

- Selectivity: The visual attention focuses on a predefined relevant visual stimulus for further processing. This is achieved through measuring the similarity with a predefined set of features. Irrelevant portion are completely ignored for further processing.

- Visual Search: Visual search focuses on the attention mechanism which is responsible for finding the target from the selective information. The capability of the visual search depends on number of destructor features in the target stimulus.

Considering these properties, in the proposed method, we use an application specific visual attention model called context saliency for measuring the saliency of the target objects in the cinematographic settings. In the next subsection, context saliency method is described in detail.

3.3 Context Saliency Map Generation

In a video frame, foreground objects are the one which are nearest to the camera and captured purposefully. Conversely, background is considered the rest of the things. A video frame might have a foreground or not. Generally, the salient regions on a video frame are the foreground objects. However, some foreground objects sometimes may not have the equal importance than the others. The context of saliency depends on the situation. For example, trees in an outdoor scenario may not have equal importance than a nearby grazing animal. In an outdoor scene, some objects are very common, such as, tree, grass, water and the sky. Conversely, the objects which are foreign in an outdoor scene has more salient property than the objects which are naturally there. Furthermore, in an indoor scene, human or other moving objects are more eye catching than the other objects in the background. In a busy road context, road signals have more saliency than the rest. Similarly, in the cinematographic context, foreground objects should have more saliency than the background portion in a movie frame. In summary, the objects which are rare yet important in a given scene should be more contextually salient than the others. Considering this in mind, we adopt an idea of computing contextual saliency from [123]. In order to compute the context saliency, first of all, the contrast saliency is computed and then the redundancy analysis is made on the contrast saliency to compute the statistical saliency. Finally, by applying geometric constraints of the frame, the context saliency is derived. Figure 3.3.1 shows a flow diagram of cinematographic shot classification and a computational flow of

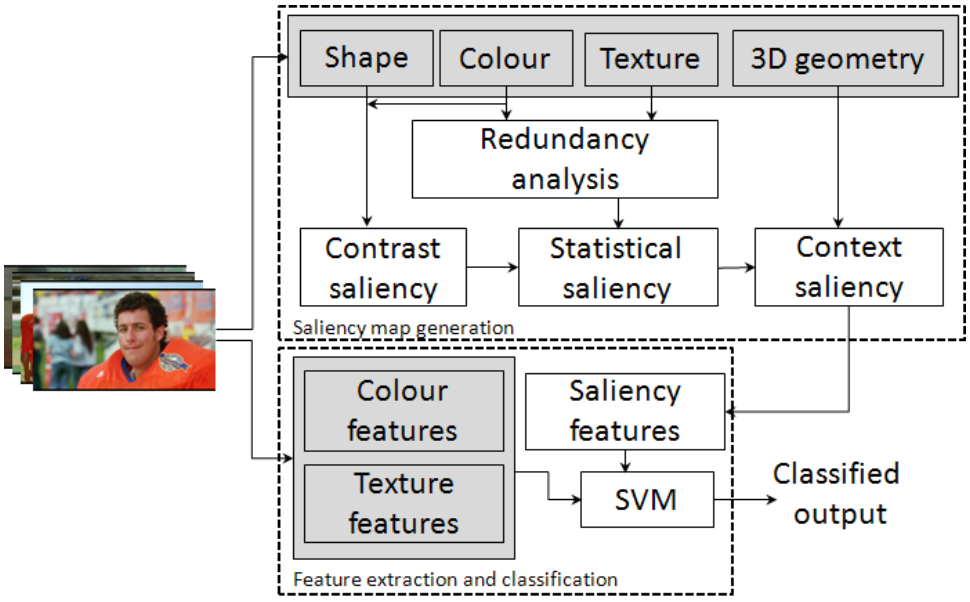


Figure 3.3.1: The framework of context saliency based video shot classification.

the context saliency map generation.

In cinematography, shots are captured to make movie characters easily attracting the attentions of audiences. There are some special cases which do not follow this. For example, establishment shots, cut shots and some close-up shots on an object when a particular part of a character or an object is important to the movie story. In most cases, characters or significant objects, which are related to movie story, belong to foreground. On the other hand, being foreground, such as trees, does not mean to be attracting audience attention or significant to movie story. In order to generate effective saliency maps for movie shot classification, we need to consider two issues.

- The salient region is more likely to be foreground rather than background.

- The salient regions should be statistically rare. Regions from frequently appeared objects and background should have low saliency.

Although some objects, such as tree, grass and mountain, are detected as foreground, these are considered as unimportant objects. Therefore, they should have low saliencies. In order to ensure that, redundancy analysis and geometric constraint are introduced. In the following, we describe the context saliency map generation technique step by step.

3.3.1 Statistical Saliency Map Generation

As shown in the framework in Figure 3.3.1, statistical saliency is generated using two inputs, namely contrast saliency and redundancy analysis. In the following, firstly, contrast saliency map computation method is discussed. Then, the redundancy analysis method is discussed to compute statistical saliency maps.

Contrast Saliency Map

Contrast is considered as an important local feature which is capable to stimulate the human visual systems. It is a commonly used feature for visual saliency based attention detection [64, 131, 135]. Human Visual Systems are highly sensitive to contrast. Ophthalmologists use letters printed on a highly contrast background in order to measure visual acuity of vision impaired patients. Thus, contrast is considered as a proven criteria in assessing human visual systems. The objects with varying surrounding contrast produces varying level of visual sensitivity. According to [136], the relationship between contrast and visual sensitivity are interrelated

which determines the level of understanding by the human visual perception.

In image processing, images are considered by three basic properties: colour, texture and shape. These properties have been applied in numerous image processing applications successfully using low level pixel analysis. Individually, each of these features cannot be used for high level understanding. Human visual systems understand high level concepts by combining different features. In fact, the contrast is the basic characteristics to be assessed from different properties to make high level concepts. The uniqueness in comparison to the surrounding environment makes an object distinct. Following this concept, finding the conspicuous locations from an input image, the contrast saliency maps are computed.

An image with $\alpha \times \beta$ pixels is considered as the *field of perception* containing $\alpha \times \beta$ *units of perception*. The contrast measure θ for the *unit of perception* at (u, v) is formally written as follows.

$$\theta_{(u,v)} = \sum_{J \in \lambda} d(I_{(u,v)}, J) \tag{3.1}$$

where $I_{(u,v)} \quad (u \in [0, \alpha], v \in [0, \beta])$ is the intensity of a pixel in a $u \times v$ region and J represents a pixel intensity of a neighbouring λ, a $u \times v$, region. The size of λ determines the *field of perception's* sensitivity. Smaller size of λ indicates higher sensitive on the field of perception. Conversely, bigger size of λ indicates lower sensitive on the field of perception. $d()$ is a suitable distance function to compute the difference between $I_{(u,v)}$ and J. In our implementation, we use 9×9 neighbourhood in RGB colour space and Euclidean distance is used for distance measurement.

In order to ensure the robustness, a multi-scale contrast saliency $\theta'_{(u,v)}$ is calcu-

lated using image intensity feature. In order to do that, contrasts in the Gaussian image pyramids are linearly combined. Formally, we write it as follows.

$$\theta'_{(u,v)} = \sum_{l=1}^{L} \theta^{l}_{(u,v)}, \tag{3.2}$$

where l indicates the l-th level of the pyramid and L indicates number of levels. After computing $\theta'_{(u,v)}$, the saliency value is normalise in the range $[0, 255]$. Figure 3.3.2 (third column form left) shows some examples of contrast saliency maps of the input frames.

Redundancy Analysis

As discussed, human faces are one of the most important objects in movie which has importance in relation to the saliency. In order to ensure the importance on human faces, a face detection module is added to find the face locations in input keyframes and boost the importance of that particular region in the contrast saliency map by a factor. We adopt a multiview face detector proposed in [137]. After detecting the faces in the input video frame, the saliency values are changed as follows.

$$\theta''_{(u,v)} = \theta'_{(u,v)} + \theta'_{(u,v)} \times c \times G(p, q, u, v, \sigma). \tag{3.3}$$

where,

$$G(p, q, u, v, \sigma) = \frac{1}{2c\sigma^2} \exp^{-\frac{(x-p)^2+(y-q)^2}{2\sigma^2}},$$

c is a constant factor, p and q represent the center of the detected face and σ represents the half of the width of the detected face. Followed by that, a redundancy analysis is applied which modifies the contrast saliency in order to compute the statistical saliency.

According to our observation, apart from the face, some common objects appear quite often in cinematography with low level of visual importance. Such objects include the sky, clouds, water, sand, ground, grass and trees. They appear as part of foreground or background. In either case, mostly they appear along with object of interest and their importance should be lower than the object of interest. Since, the visual importance of these objects should be lower than the objects of interest, we take an strategy to identify those locations using a supervised learning technique to reduce the contrast saliency values. In order to do that, 50 keyframes are collected from cinematographic shots consisting indoor and outdoor scenes. From the keyframes, foreground and background regions are identified and labelled manually. In order to do that, if there is any person present then that person is considered as foreground. The rest of the objects are considered as background. Then, each of the keyframes is divided into 9×9 pixel patches. For each of the patches, the colour histogram and the gray level co-occurrence matrix are computed. Then, using the extracted features, K-mean clustering is applied on all the patches and 7 representative clusters are identified. The representative 7 clusters represent the sky, clouds, water, sand, ground, grass and trees roughly. After that, using these representative clusters, the contrast saliency maps are modified by using a 9×9 sliding window. Mainly, we compute the density of information by

using the distance between the sample patches with current patch and neighbourhood of the current patch. In order to do that, we consider the image layout and the distance to the sample patches. Statistical saliency is described as follows.

$$\psi_i = \theta_i^{''} \times -\log_2 \frac{\psi_{i-1}}{\min\{Dis(h_i, h_s)\}}, \tag{3.4}$$

where i is the index of the continuous patches. $Dis(h_r, h_s)$ is bin by bin colour histogram distance between current patch i and the representative sample patches. ψ_i is then normalised as follows.

$$\psi_i^{'} = \frac{\psi_i}{max(\psi_i)} \tag{3.5}$$

where $max(\psi_i)$ is the maximum value of ψ_i. In order to reduce the redundancy and blob noise, the patches with low information density are removed from statistical saliency maps.

Extracting 3D Geometric Information

Finding the orientation of the image is useful for generating the depth concept from a single image. We adopt the geometric context extraction method from a single image [138]. It estimates the coarse geometric properties of a scene by learning appearance-based models of geometric classes with a multiple hypothesis framework. In order to determine the orientation, it is important to use all available information. For example location, texture gradients, shading and vanishing points. Most of these information can only be possible to extract when some

important knowledge is available. The solution proposed in [138] builds structural knowledge by using superpixels. The first step is to apply an efficient graph based image segmentation method to obtain a set of *superpixels* proposed in [139]. Each superpixel has a single label. It is possible to measure some basic first order statistics (e.g., colour and texture) by using the superpixels and their corresponding boundaries. Then, superpixel label confidences, weighted by the homogeneity likelihood, are determined by averaging the confidence in each geometric label of the corresponding regions as follows.

$$\gamma(y_i = v|x) = \sum_{j}^{n_h} P(y_j = v|x, h_{ji}) P(h_{ji}|x) \tag{3.6}$$

where γ is the label confidence, y_i is the superpixel label, v is a possible label value, x is the image data, n_h is the number of hypotheses and h_{ji} defines the region containing the ith superpixel for the jth hypothesis with the region label y_j.

With statistical saliency $\psi'_{(x,y)}$ and geometric constraint $\gamma_{(x,y)}$, context saliency ω is modelled by Bayesian framework using the maximum a posteriori (MAP) criterion described in the following subsection.

3.3.2 Context Saliency Map Generation

With the computed statistical saliency $\psi'_{(x,y)}$ and geometric constraint $\gamma_{(x,y)}$, the expected desired feature of context saliency ω is modelled by Bayesian framework by using the maximum a posteriori (MAP) criterion. The most likely context saliency ω is estimated, given ψ' and γ. This is expressed as maximisation of a proba-

bility distribution over a sum of log likelihood. Formally, we write it as follows.

$$\arg\max P(\omega_{x,y}|\psi'_{x,y}, \gamma_{x,y}) = \arg\max L(\psi'_{x,y}|\omega_{x,y}) + \log P(\gamma_{x,y}|\omega_{x,y})P(\omega_{x,y}). \quad (3.7)$$

where $L()$ is the likelihood function. We assume $P(\omega_{x,y}) = 1$. The computed map $\Omega \in \{\omega_{x,y}\}$ is the context saliency map. Figure 3.3.2 (right column) shows the output of context saliency maps for commonly appeared cinematographic shot types.

3.3.3 Image Intensity Histogram

The image intensity histogram is used as another feature in the proposed method. In order to compute this feature, we compute the weighted intensity of the image pixels $P = \{p_{x,y} : x \in a, y \in \beta\}$ using the following equation.

$$p_{x,y} = 0.299 \times r_{x,y} + 0.587 \times g_{x,y} + 0.114 \times b_{x,y}. \quad (3.8)$$

where $r_{x,y}$, $g_{x,y}$ and $b_{x,y}$ represent red, green and blue channels of the pixel at (x, y) respectively. Then, the histogram of the gray level image is computed using the image pixels. Formally, the histogram is described as follows.

$$H(P) = \Big(h_1(c), h_2(c), \ldots, h_n(c)\Big) \quad (3.9)$$

Figure 3.3.2: Examples of most commonly appeared shot types in movies, corresponding contrast saliency maps (third column from left)and context saliency maps (right column) using proposed method.

where h_i represents a histogram bin index, $h_i(c)$ represents the bin count and n represents the number of bins. In order to represent the histogram in a compact manner, we quantise the pixel intensity value to form a 32 bin histogram.

3.3.4 Pixel Correlation

The inter-pixel relationship is used to represent the textureness of a keyframe. To measure the texture, a pixel correlation is computed. Equation 3.10 is used to measure the pixel correlation. We compute the correlation $c_{x,y}$ by aligning a box filter [140] with a pixel at (x,y). Then, by applying the box filter on the pixels, we sum up the total values to measure the correlation coefficients. We write it as follows.

$$c_{x,y} = \sum_{j=-N}^{N} \sum_{i=-N}^{N} I(x+i, y+j) * f(i,j) \tag{3.10}$$

where $f(i,j)$ is a box filter of size $(2 \times N + 1) \times (2 \times N + 1)$. $I(x,y)$ represents the pixel intensity at the corresponding location. After computing pixel correlation, we compute the mean of the pixel correlation as a feature.

3.3.5 Entropy

Local information content within a keyframe is another feature. To compute the entropy, the keyframe pixel histogram is computed. Using the histogram bins, local entropy is computed using the following equation.

$$E = -\sum_{i=1}^{K} h_i(c) \log(h_i(c)) \tag{3.11}$$

where i is index, K is the number of bins and $h_i(c)$ is the total number of counts in the histogram at bin index i.

3.4 Feature Extraction for Cinematographic Shot Classification

We extract 80 dimensional features from both context saliency maps and the original keyframes to describe global and local characteristics. Firstly, image intensity histogram is calculated with 32 bins from the original keyframes as global features. As local features, firstly, features representing the distribution of salient regions are extracted from context saliency maps. Saliency maps are divided into $16\,(=4\times4)$ equally sized local regions. The saliency value is calculated for each local region to get 16 dimensions of features. For each of the local regions, total local context saliency magnitude $\Omega_{(p,q)}$ is computed by summing all pixel's saliency values within that local region. Equation (3.12) expresses the local context saliency feature formally.

$$\Omega_{p,q} = \sum_{x\in p, y\in q} \omega_{x,y} \tag{3.12}$$

where $p \in \{1,2,3,4\}$ and $q \in \{1,2,3,4\}$ are local region indices and $\omega_{x,y}$ represents context saliency of a pixel at (x,y). After this, each local context saliency magnitude is normalised using the total saliency magnitude of the keyframe. We

write it as follows.

$$\bar{\Omega}_{p,q} = \frac{\Omega_{p,q}}{\sum_{x \in \alpha, y \in \beta} \omega_{x,y}} \tag{3.13}$$

where α and β represent width and height of the image respectively.

Then, correlation and entropy features are computed. For both cases, input keyframes are divided into $16 = (4 \times 4)$ equal sized regions. For each local region (p, q), we compute the normalised mean local correlation coefficient values as local features.

$$C_{p,q} = \frac{\sum_{x \in p, y \in q} c_{x,y}}{\sum_{i \in \alpha, j \in beta} c_{i,j}} \tag{3.14}$$

where $p \in \{1, 2, 3, 4\}$ and $q \in \{1, 2, 3, 4\}$ indicate local regions and α and β indicate image height and width respectively. Similarly, for each local region at (p, q), the normalised local entropy features are computed as follows.

$$E_{p,q} = \frac{\sum_{x \in p, y \in q} E_{x,y}}{\sum_{i \in \alpha, j \in \beta} E_{i,j}} \tag{3.15}$$

where $p \in \{1, 2, 3, 4\}$ and $q \in \{1, 2, 3, 4\}$ indicate local regions and α and β indicate image height and width respectively.

3.5 Experimental Results

To show the performance of the proposed features in cinematographic shot classification, we present the experimental results in this section. In order to evaluate the

performance, we have created a cinematographic keyframe dataset. The extracted keyframes are taken from a wide range of movie. The movies which are used to create this dataset are: The Proposal, The Terminal, The Kids Are All Right, Little Miss Sunshine, Prison Break, The Beautiful Mind, Lord of the Rings and Mission Impossible II. The keyframes of the created dataset are divided into two parts: training and testing set. For training the classifiers, we use the training set and the classification performance is evaluated using the testing set. The performances are measured using precision rates, recall rates and f_1-scores. In the following subsections, we describe the detail of the experiment and the evaluation procedure.

3.5.1 Dataset Preparation and Feature Extraction

To evaluate the classification performance of the proposed context saliency and other related features, the created dataset of 3206 keyframes are extracted from the first frames of individual shots and are manually classified into six cinematographic shot classes based on camera and objects positions (described in Chapter 1). The classes are: 1) Close-up (CU) shot, 2) Over the shoulder (OTS) shot, 3) Medium close-up (MCU) shot, 4) Medium (M) shot, 5) Cut (C) shot, and 6) Long (L) shot. For training, we use one third of shots from each class. The detail breakdown of the keyframe dataset is given in Table 3.5.1.

Table 3.5.1: Detailed breakdown of the testing data for context saliency based shot classification.

	CU	OTS	MCU	M	C	L
no. of shots	1630	210	389	339	232	406

We extract an 80 dimension feature vector representing local and global features from the created dataset. Classifiers are trained using the features extracted from the training set. Then, the performance is measured using the testing set. In the following, we discuss about the SVM classification technique used in the proposed method.

3.5.2 SVM Classification

Vapnik *et al.* introduced the famous support vector machine (SVM) technique for binary classifications by using the principal of statistical learning theory [141]. To generalise the proposed technique, many research efforts have been accomplished to extend it to a multi-class SVM. For a k class classification problem, three major approaches can be applied, namely One-Against-One, One-Against-All and DAGSVM [142].

In this chapter, we use One-Against-All SVM technique for keyframe classification. To summarise the One-Against-All multi-class SVM, let us assume that we have n training data in d dimensional space belonging to c classes $\{\mathbf{x}^i, y^i\}, \quad \mathbf{x}^i \in \mathbf{R}^d, \quad i = 1, \ldots, n, \quad y^i \in \{1, \ldots, c\}$. This approach constructs c classifiers using the training data. Each classifier is obtained by using the training data of the corresponding two classes. Training samples belong to class i are labelled with a positive label and the reset of the training samples are labelled with negative label.

This binary classification problem is formally written as:

$$\begin{aligned} &\min_{w_i, b_i, \xi_i} \quad \left(\frac{1}{2} (w_i)^T w_{ij} + C \sum_t \xi_i^t (w_i)^T \right) \\ &(w_i)^T \varphi(x_j) + b_i \geq 1 - \xi_j^i, \text{if } y^j = i \\ &(w_i)^T \varphi(x_j) + b_i \leq -1 + \xi_j^i, \text{if } y^j \neq i \end{aligned} \tag{3.16}$$

where ξ_{ij}^t is a non-negative slack variable, $\varphi(x^i)$ is a function to map x^i into a higher dimensional space and C is the penalty parameter. By minimising $\frac{1}{2}(w_{ij})^T w_{ij}$, we want to maximise the margin, $\frac{2}{||w_{ij}||}$, between class i and class j. The penalty term $C \sum_t \xi_{ij}^t$ is used to reduce the number of training errors for linearly non-separable cases. The goal is to find an optimal separating hyperplane by obtaining a balance between the regularisation term $\frac{2}{||w_{ij}||}$ and the training errors. To improve the separability, the data are mapped into a higher dimensional dot product space using the function φ. If the dot product space is expressed by $K(x^i, x^j) = \varphi(x^i) \cdot \varphi(x^j)$, then $K(x^i, x^j)$ is called the kernel function. The kernel used must meet Mercer's condition which is described in [141]. The accuracy of SVM classification depends on the values of two parameters C and γ. Careful selection of these two parameters is important. Otherwise the classifier may perform poorly in the testing phase. A cross-validation approach is commonly used to determine the best parameters. We find the best penalty parameter C from the range $\{2^{-5}, 2^{-4}, \ldots, 2^{10}\}$ and width control parameter γ from the range $\{2^{-10}, 2^{-1}, \ldots, 2^{5}\}$.

3.5.3 Evaluation

The performance of the proposed features is evaluated by using SVM classifier. For the created keyframe dataset, the effectiveness is shown by using confusion matrix and by computing recall rates, precision rates, and f_1-score. The precision rates and recall rates are measured as follows.

$$precision = \frac{tp}{tp + fp}, \tag{3.17}$$

$$recall = \frac{tp}{tp + fn} \tag{3.18}$$

where tp is true positive, fp is false positive and fn is false negative. f_1 scores are measured as follows.

$$f_1 score = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.19}$$

Table 3.5.2 shows the confusion matrix of the shot classification performance. The recall rates and precision rates are reported in Table 3.5.3 and Table 3.5.4 respectively. As shown in the tables, the correct classification performance of CU, M and L is reasonably high. However, the performance of OTS, MCU and C is considered to be poor. Although the proposed features could not perform evenly for all shot classes, the performance of the proposed method is considered as acceptable. Because, the higher performing classes consist of 74% of our keyframe dataset. Some video shots, such as cut shots and medium shots appear relatively

less frequently. The number of shots belonging to each class is relatively unbalanced and it should be noted that the ratio of the samples within a dataset heavily affects the degree of classification accuracy. Another possible issue which affects classification is data labeling. While manually labeling the data, we find it is not easy to distinguish close-up from medium shot for some cases. In movies, these two shots are actually framed similarly. Mislabelling of close-up shot and medium shot somehow affects the classification of these two shot types.

Table 3.5.2: Confusion matrix for shot classification using the keyframe dataset.

	CU	OTS	MCU	M	C	L
CU	77.0	7.7	6.5	3.7	1.8	3.4
OTS	26.2	52.4	12.4	7.1	1.4	0.5
MCU	15.6	6.4	43.7	8.0	1.0	2.6
M	7.4	4.4	17.7	61.9	6.8	1.8
C	19.0	7.8	19.0	6.9	39.2	8.2
L	6.2	5.2	2.7	8.9	4.2	72.9

Table 3.5.3: Recall rates for shot classification.

shot type	CU	OTS	MCU	M	C	L
Recall	77.0	52.4	43.7	61.9	39.2	72.9

Table 3.5.4: Precision rates for shot classification.

shot type	CU	OTS	MCU	M	C	L
Precision	86.4	35.0	40.8	57.1	54.5	76.5

In order to justify that context saliency plays an important role in classification, we classify close-up shots using three different sets of features. Figure 3.5.1, shows

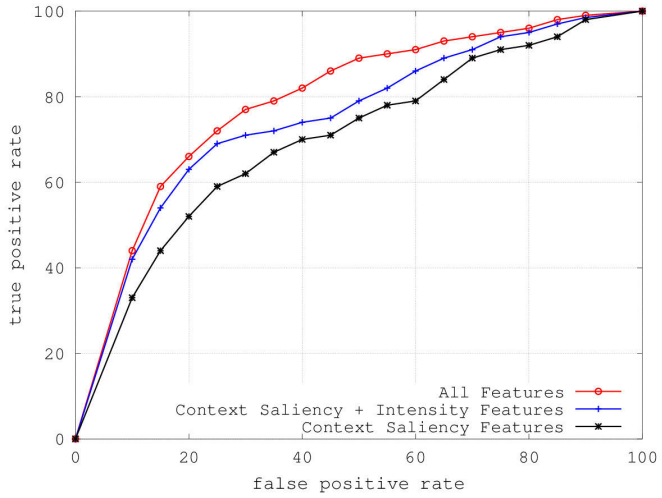


Figure 3.5.1: ROC curves for Close up shot classification using different feature sets.

the ROC curves of close up shot classification using different feature sets. As it can be seen, context saliency features have a significant contribution in classifying close up shots. Moreover, inclusion of additional features boost the classification performances. From the figure, it is obvious that saliency features have significant importance in shot classification. The f_1 score of CU classification using only context saliency features is 73.25 whereas the f_1 score of CU classification using all features is 81.4. Figure 3.5.2, shows that the f_1 scores of CU, M, and L shots are higher than the rest shot types. These three are the top three popular shots in our dataset while other shots consist of only 25% of the total shots.

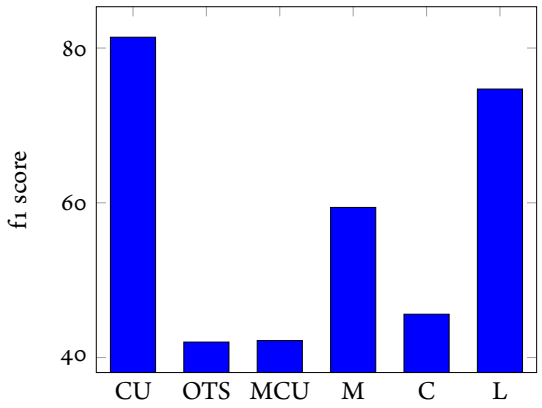


Figure 3.5.2: f_1 scores of different classes using keyframe dataset.

3.6 Summarisation

Cinematographic shot classification is a vital and challenging task. In this chapter, we have proposed a context saliency based cinematographic shot classification method. The context saliency has been produced by removing redundancies with low information densities from the contrast saliency and incorporating geometry constrains. Compared to the traditional saliency map, the context saliency represents the visual attention distributed in a video frame. Through experiments, the context saliency has been proved to be significant for shot classification although unbalanced cinematographic shots have somehow affected classification results.

Be the change that you wish to see in the world.

Mahatma Gandhi

4

Hierarchical Approach to Cinematographic Shot Classification

4.1 Introduction

In this chapter, we propose a hierarchical approach to cinematographic shot classification by introducing a set of cinematographic domain specific features. The domain specific features are used in two phases to get better classification results than the results achieved in Chapter 3. First of all, shots are classified into one of the top three depth based shot classes, they are named close distance shot (CDS), medium distance shot (MDS) and long distance shot (MDS). Then, each shot is further classified into one of the pertaining classes. For classification, we use SVM classifiers. The contributions of this chapter are as follows.

- We analyse the cinematographic domain shots and introduce a set of movie domain specific features for cinematographic shot classification.
- We propose a novel hierarchical approach for cinematographic shot classification.

In Chapter 3 and in one of our previous works [1], we have classified cinematographic shots based on object and camera position using cinematographic context saliency maps. The performance of the introduced features is promising although there is still a scope to improve the results. To obtain better classification results, we dig the problem further for a better performance. In Section 4.2, the detail of the proposed framework is discussed.

4.2 Framework for the Hierarchical Approach

In our proposed hierarchical approach to cinematographic shot classification framework, there are two distinct levels of classification involved. The first level of classification involves classifying shots into one of the top three depth based classes, as defined CDS, MDS, and LDS. If a shot belongs to CDS or MDS, then the shot has to go through one more level of classification. Figure 4.2.1 shows the overall framework of the proposed method. In our previous work [143], we use a set of movie domain specific features for cinematographic shot classification. The following subsections describe the details of our proposed features and the classification techniques.

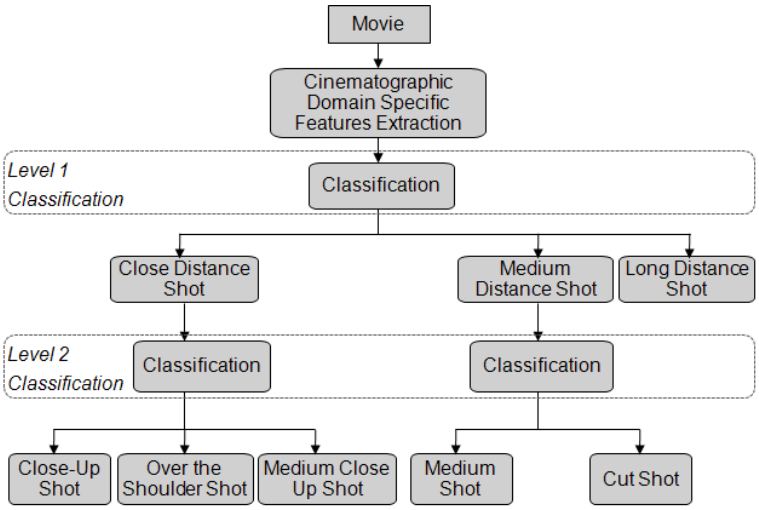


Figure 4.2.1: Flow diagram of hierarchical cinematographic shot classification.

4.3 Cinematographic Domain Specific Features

Cinematographic domain video shots are significantly different from the other types of video shots. Some of the characteristics of cinematographic shots are distinct and we exploit those characteristics for the purpose of distinctly characterising cinematographic shots. In this section, we introduce the proposed cinematographic domain features. As mentioned in the previous chapter, irrespective to any movie genres, actors take important roles in movies and facial appearance is one of the most common characteristics in movies. Apart from that, several other characteristics can be derived. The video shots in movie have some distinctive features. In producing quality shots, the use of the modern cameras also plays a big role. Moreover, well-defined spatial compositional rules are applied in capturing such shots. Considering these issues, a set of movie domain features are extracted from the keyframes. In the following, we discuss about the proposed features in

detail. The proposed features have been published in one of our previous work in [143].

4.3.1 Weighted Hue Histogram

Colour is a widely used important feature to describe an image. In HSI colour space, the hue represents the colour information while the saturation represents the purity of the colour. Due to changing intensity of lights, the saturation may vary significantly. In a movie setting, by using hue analysis, we can get the rough idea of the scene depth from the objects present in a scene. For example, in a long outdoor shot, a significant portion of the image usually contains the blue sky. Similarly, in a close-up shot, a significant portion of the shot usually contains human skin colour. By analysing hue histogram, the dominant colour of a frame is identified and can be used as a describing feature of a cinematographic keyframes. Figure 4.3.1 shows hue histograms of different types of example shots. As it can be seen, the corresponding histogram of each of the shot represents the dominant hue of each keyframe. In the left keyframe, face colour is the most dominant hue and the corresponding histogram represents red as the most dominant hue. Similarly, in a forest shot, green tone is more dominant. In order to compute the hue histogram, firstly, the RGB colour channels are used to compute hue value Φ for the pixel at (x, y), denoted by $\varphi_{x,y}$ where $\Phi \in \{\varphi_{x,y}\}$. The equation to measure hue value is as

follows.

$$\varphi_{(x,y)} = \begin{cases} \cos^{-1}\left\{ \frac{0.5\times[(r_{(x,y)}-g_{(x,y)})+(r_{(x,y)}-b_{(x,y)})]}{[(r_{(x,y)}-g_{(x,y)})^2+(r_{(x,y)}-b_{(x,y)})(g_{(x,y)}-b_{(x,y)})]^{\frac{1}{2}}} \right\} & \text{if } b_{xy} \leq g_{xy} \\ 2\pi - \cos^{-1}\left\{ \frac{0.5\times[(r_{(x,y)}-g_{(x,y)})+(r_{(x,y)}-b_{(x,y)})]}{[(r_{(x,y)}-g_{(x,y)})^2+(r_{(x,y)}-b_{(x,y)})(g_{(x,y)}-b_{(x,y)})]^{\frac{1}{2}}} \right\} & \text{otherwise} \end{cases} \tag{4.1}$$

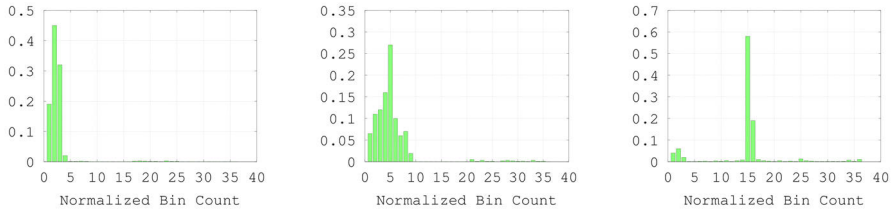
where $r_{(x,y)}$, $g_{(x,y)}$ and $b_{(x,y)}$ represent *red*, *green* and *blue* channels of the corresponding location in the RGB colour space respectively. The hue of each pixel $\varphi_{(x,y)}$ is computed in the range of $0 \leq \varphi_{(x,y)} < 2\pi$. Computing a bin for each degree is too fine and unnecessary for describing the colour. Instead of that, we can quantise the hue into a certain number of levels for a compact representation. Using the computed hue values, a hue histogram H_φ of an input keyframe is formulated as follows.

$$H_\varphi(c) = \left(\varphi^1(c), \varphi^2(c), \ldots, \varphi^n(c)\right) \tag{4.2}$$

where c represents the count of the i-th hue bin and n represents total number of bins.

Since the foreground content has strong influence in determining the cinematographic shot type, we develop a strategy to incorporate the contextual salient region information in the hue histogram. The proposed scheme is called the weighted hue histogram. Our intention is to emphasize on the foreground pixels. In order to do that, we introduce a novel weighted hue histogram scheme. Firstly, the con-

(a) sample video frames



(b) corresponding hue histogram

Figure 4.3.1: Hue histogram analysis of three different types (CDS, MDS, LDS) of shots. For CDS, as the frame mostly comprises with human face, the hue is mostly dominated by red-yellow tone. For the MDS and LDS hue represents most dominated colours.

text saliency map is thresholded using an entropy thresholding [144] method. The pixel of binarise map is defined by $I_{(x,y)} \in \{0, 1\}$. Then, by using the thresholded context saliency map, we compute the mean μ_φ and standard deviation σ_φ of the hue distribution. Formally, we write it as follows.

$$\mu_\varphi = \frac{\sum_x \sum_y \varphi_{(x,y)} \times I_{(x,y)}}{\sum_x \sum_y 1 \times I_{xy}}, \tag{4.3}$$

$$\sigma_\varphi = \sqrt{\frac{(\sum_x \sum_y \varphi_{(x,y)} - \mu_\varphi)^2}{\sum_x \sum_y 1 \times I_{(x,y)} - 1}}. \tag{4.4}$$

Using the mean and standard deviation, we compute weighted hue histogram of the keyframe. The intention is to identify the variation in the salient portion of

the frames. If the variation is low, then it indicates that the salient region hue contents are from homogeneous pixels. Therefore, only the corresponding pixels have more importance in the hue histogram. We define the weighted hue histogram as follows:

$$H_{\varphi}^{\omega}(c) = \Big(f(\varphi^1(c)), f(\varphi^2(c)), \ldots, f(\varphi^n(c))\Big) \tag{4.5}$$

where,

$$f(\varphi^i(c)) = \varphi^i(c) \times exp\left(-\frac{\varphi^i - \mu_\varphi}{2\sigma_\varphi^2}\right).$$

where $f(\varphi^i(c))$ indicated the bin count of the i-th bin of the histogram.

4.3.2 Edge Features

An image is constructed by transforming the lights coming from 3D world in the form of 2D and projecting onto a 2D geometric surface. The transformation of the lights can be controlled, and in this regard, today's state-of-the-art movie cameras play a big role. Cameramen capture a cinematographic shot by controlling the camera to let the lights in from the target objects effectively. Movie cameras usually capture 24 frames per second and thus, the exposure time is fixed to 1/24 sec. In order to control lights, cameramen use two techniques.

- Using light filter; and
- Using camera aperture.

Exposure is a big concern in quality photography. In extremely bright light condition, light filters are useful to control incoming lights which is used to filter out

unnecessary incoming lights. Another way of controlling incoming light is to use the camera aperture. The size of aperture determines the amount of lights permitted to enter through the lens. The size of the aperture is measured in *f*-numbers and indicates the diameter of lens opening. More opening permits more lights to be entered in the camera. Apart from light controlling, by using the aperture, cinematographers controls depth of fields on the image plane. For creating a shallow focus (or small depth of field), a large aperture is used (*e.g.* f/1.4, f/2 and f/2.8). Similarly, for creating a deep focus, a small aperture is used (*e.g.* f/22 and f/29). The law which describes the characteristics of the geometrical transformation of the optics is as follows.

$$\frac{1}{f} = \frac{i}{v} + \frac{1}{d}, \tag{4.6}$$

where f is the focal length of the lens, v is the distance between image plane and the lens axis, and d is the object distance from the lens axis. Figure 4.3.2 shows lens geometry of a basic camera model. As shown, the lights from the focused object meet on the image plane. However, lights from other objects meet before they reach the image plane, and create blurry patches on the image plane.

In a movie camera, the depth of field is determined by the distance of object (d), lens focal length (f) and the changeable aperture size (f-number). Use of aperture restricts the distance of the image plane to the lens focal length. Because of that, focused objects have sharp appearance on the image plane. However, lights which are not from the focused objects, do not converge on the image plane. As a re-

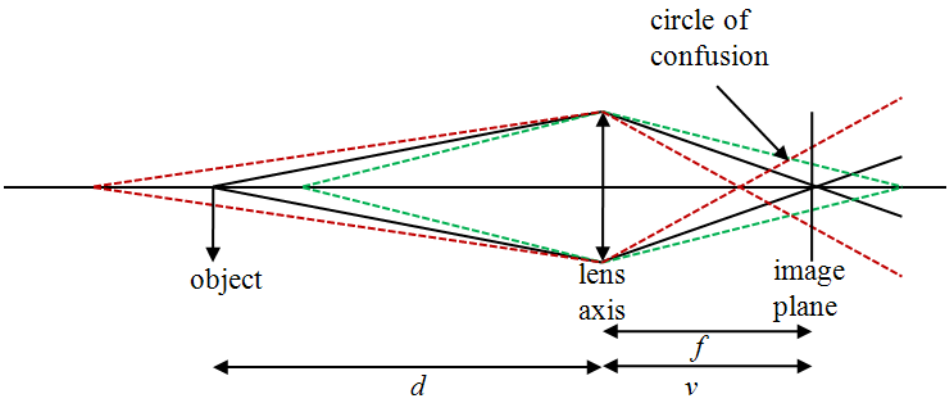


Figure 4.3.2: Lens Geometry: light from each point of focused objects meets on the image plane, while out of focus objects create blurring effect by creating circles of confusion.

Figure 4.3.3: Shallow focus: characters are sharply in focus while the backgrounds are blurred.

sult, circles of confusion appear. Circles of confusion create blurry patches in the images.

In the case of capturing a close-up shots, particularly in cinematography, cinematographers use shallow focuses, which invoke low depth of fields. As a result, subjects appear sharply in focus while the backgrounds appear out of focus. Figure 4.3.3 shows some examples of close up shots. As shown, shallow depth of field creates sharp focus on the subject while background appears as smooth patches without much detailed information.

Figure 4.3.4: Deep focus: everything in the frames are sharply in focus.

In contrast, long shots usually provide every detail of the contents. To ensure that, a deep focus is used. In order to increase the depth of field in capturing a deep focus, a small aperture is used. This mechanism ensures lights coming from different objects of different depths to correctly converge on the image plane. Therefore, sharp shots are produced. Figure 4.3.4 shows some examples of long shots with deep focuses.

Based on the above discussion, it is obvious that different camera settings create a changing level of edge energy. We summarise that there is a strong correlation between edge energy and the depth of the field. Therefore, the edge energy is considered as a strong feature in discriminating cinematographic shots. In order to measure edge energy, we compute Haar wavelet coefficients. Wavelet is an excellent tool to determine sharpness of the edge. The edge energy computation is performed in two steps. Firstly, a keyframe is decomposed up to n levels using Haar wavelet. Then, edge energy is computed as follows.

$$e_i = \sqrt{lh_i^2 + hl_i^2 + hh_i^2} \tag{4.7}$$

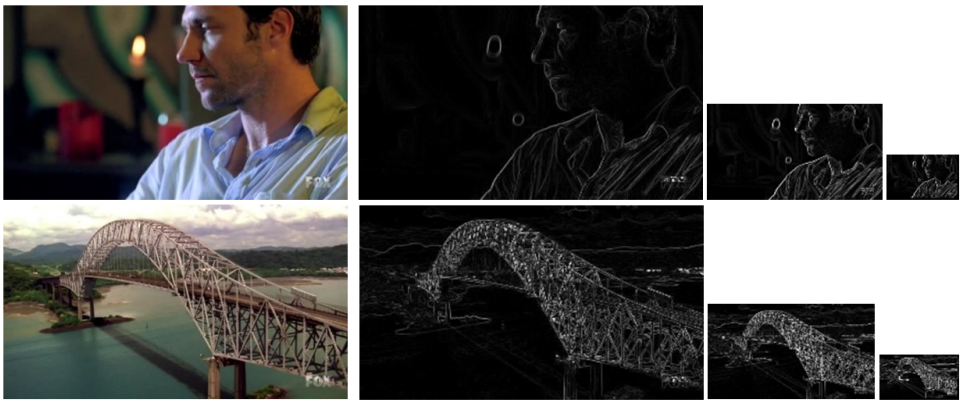


Figure 4.3.5: Computed wavelet edges up to three level using Haar wavelet transformation.

where i represents the decomposition level. lh, hl and hh represent low-high, high-low and high-high sub-band respectively. Figure 4.3.5 shows examples of edge decomposition up to 3 levels.

4.3.3 Region with Skin Colour

Movie is mainly all about human activities, human facial expressions and body languages. Based on the requirements, the distance from a human body to a camera varies significantly. In order to measure the depth of a scene, the area of skin region provides a clue of distance of objects from the camera. In order to measure the area of skin colour region, we need to segment skin region pixels from an input frame. In the literature, a wide number of skin colour detection methods are proposed. A good survey of these method can be found in [145]. Skin segmentation is in movies is different than the normal lighting settings. In movies, due to the use of different light sources, skin colour varies significantly for different indoor and

outdoor shootings. Moreover, different lighting creates different skin tones of the same person. Therefore, we need to consider this issue in order to segment skin region robustly. In the proposed method, we use a simple but effective threshold based technique to segment regions with skin colour. Firstly, a multiview face detector [137] is used to identify the face regions within the keyframes. Then, using the detected face regions, we measure the mean $\mu_s = \{\mu_r, \mu_g, \mu_b\}$ and covariance matrix Σ_s of facial regions. Then, using these two parameters, each pixel of the corresponding keyframe is labelled as a skin or non-skin pixels as follows.

$$S(p_{x,y}) = \begin{cases} 1 & \text{if } dist(p_{x,y}, \mu_s, \Sigma_s) < \tau \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

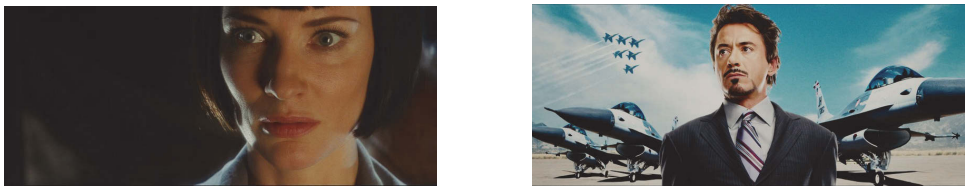
where,

$$dist(p_{x,y}, \mu_s, \Sigma_s) = (p_{x,y} - \mu_s)^T \Sigma_s^{-1} (p_{x,y} - \mu_s),$$

$p_{x,y}$ represents pixel colour in RGB colour space and τ is a predefined threshold. If a keyframe does not contain any face, then the previous keyframe's parameters are used in segmenting skin regions. Figure 4.3.6 shows examples of the segmentation results of skin pixels.

4.3.4 Other Features

The context saliency map, pixel correlation and entropy, which have been described in Chapter 3, are also considered as cinematographic domain features and hence used in the proposed framework in this chapter.



(a) sample cinematographic frames

(b) corresponding segmented skin pixels

Figure 4.3.6: Examples of skin colour segmentation.

4.4 Proposed Hierarchical Shot Classification

In Chapter 3, cinematographic shots are classified into 6 classes by using context saliency maps, colour intensity and pixel correlation features. In this section, we propose a hierarchical shot classification technique with an intention to improve the classification performance. First of all, the cinematographic shots are categorised into three top level classes based on the distance from a camera. We have named them, CDS, MDS and LDS. Close-up shots, over-the-shoulder shots and medium close-up shots belong to CDS. Medium shots and cut-shots belong to MDS. There is no classification required in the LDS and a classified LDS is directly considered as a long shot. As mentioned, there are two phases in our proposed technique. The first phase, *level 1 classification*, involves classifying a shot into one of the top three classes: CDS, MDS and LDS. After that, if the shot belongs to CDS

Figure 4.4.1: Examples of the rule of thirds.

or MDS then the shot has to go through the second phase of classification called *level 2 classification*. The following subsections describes the details of methods for feature extraction at level 1 and level 2.

4.4.1 Features for Level 1 Classification

In order to classify a cinematographic shots into one of the top three classes, we extract local and global features using the feature extraction methods discussed in section 4.3. The local features include context saliency map, region with skin colour, pixel correlation and entropy. The global features include weighted hue histogram, edge features and number of faces in the a keyframe.

Local Features Computation

In order to compute the local features, a popular photographic rule, rule of thirds (ROT) is used. The ROT is a widely practised composition rule in the professional photography. This rule describes an effective way of positioning the objects of interest within the frame. In order to do that, a frame is imaginatively divided into a 3×3 grid. The intersecting points of the imaginary lines are considered

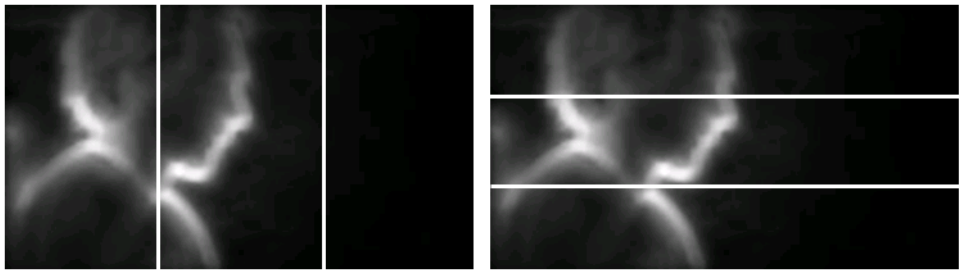


Figure 4.4.2: Saliency map is segmented in vertical and horizontal directions to compute saliency features.

as golden locations, where the important objects are placed. For example, the facial region of a close-up shot is intentionally placed in one of the golden locations. Moreover, each imaginary section contains information to best describe the scene. For example, in the case of a long shot in a outdoor scene, horizon line is placed roughly either along the top imaginary line or along the bottom imaginary line. If the horizon is placed along the lower imaginary line, then it is obvious that the photo mainly tells a story about the sky. If the horizon is placed along the upper imaginary line, then the photo obviously tells a story about the ground or water. Figure 4.4.1 shows examples of ROT.

The proposed local features are computed, considering that the ROT as an important composition rule in professional photography. At the beginning, local features are extracted from the context saliency maps. Firstly, each context saliency map is segmented into 3 horizontal and 3 vertical regions. Figure 4.4.2 shows an example of segmenting a context saliency map in horizontal and vertical directions. Then, for each of the horizontal and vertical segments, the total local context saliency magnitude Ω_i^o is computed by summing all pixel's saliency values within

that local region. Equation (4.9) expresses the local context saliency feature formally.

$$\Omega_i^o = \sum_{x \in i, y \in i} \omega_{x,y} \tag{4.9}$$

where o is orientation of segmentation, which can be h or v representing horizontal and vertical directions respectively. $i \in \{i, 2, 3\}$ indicates the index of the corresponding segment. $\omega_{x,y}$ represents the context saliency of the pixel at (x, y). After this, each of local context saliency magnitude is normalised using the total saliency magnitude of the keyframe. We write it as follows.

$$\bar{\Omega}_i^o = \frac{\Omega_i^o}{\sum_{x \in a, y \in \beta} \omega(x, y)} \tag{4.10}$$

where a and β represent width and height of the image respectively.

In a similar way, the region containing skin pixels is segmented into 3 horizontal and 3 vertical regions. For each of the horizontal or vertical segment, the total number of skin pixels is counted. Then, for each local region's pixel count is normalised using total number of pixels of the keyframe. We write it as follows.

$$S_i^o = \frac{\sum_{x \in i, y \in i} S(p_{x,y})}{\sum_{x \in a, y \in \beta} S(p_{x,y})} \tag{4.11}$$

The pixel correlation and entropy features are computed by segmenting the keyframe into 9 (3×3) equal sized local regions. For each of the local regions, pixel correlations C_i^h, C_i^v and entropies E_i^h, E_i^v are computed. Then, each of computed corrections and entropies are normalised using the total correlation and entropy

respectively. We write it as follows.

$$C_i^o = \frac{\sum_{x \in i, y \in i} c_{x,y}}{\sum_{i \in a, j \in beta} c_{i,j}} \tag{4.12}$$

where $i \in \{1, 2, 3\}$ indicates a local region and α and β indicate the image height and width respectively. Similarly, for the local region at (p, q), the normalised local entropy features are computed as follows.

$$E_i^o = \frac{\sum_{x \in i, y \in i} E_{x,y}}{\sum_{i \in a, j \in \beta} E_{i,j}} \tag{4.13}$$

where $i \in \{1, 2, 3\}$ indicates a local region and α and β indicate the image height and width respectively.

Global Features Computation

In order to represent the hue information, selecting quantisation is important. Having too many bins in the histogram creates a high dimensionality in the feature space. To represent the primary colours, we use 12 bins in the proposed weighted hue histogram. We count the bins in such a way that each bin describes one of the distinguishable colours from the hue circle. Figure 4.4.3 shows an example of hue circle. In the hue circle, each 60° interval (e.g. 0, 60 and 120) represents a pure colour. Between two pure colours, another colour is created as a result of the mixture of the colours (e.g. orange colour is appeared as a mixture of red and yellow). Between 45° and 74°, it represents a yellow colour. Using this observation, we quantise the hue circle for each 30°. We compute a 12 bin weighted hue

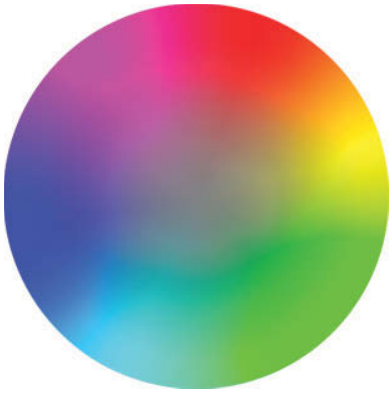


Figure 4.4.3: Exmple of the hue circle.

histogram using Equation 4.3.1. The weighted hue histogram is normalised for a uniform representation using the following equation.

$$\bar{H}^{\omega}_{\varphi}(c) = \left(\frac{f(\varphi^1(c))}{C}, \frac{f(\varphi^2(c))}{C}, \ldots, \frac{f(\varphi^n(c))}{C}\right) \tag{4.14}$$

where $C = \sum_i^n \theta^i(c)$. Equation (4.14) is used as the colour feature to represent the colour distribution in the keyframe.

Edge energy is computed as another global feature. In order to identify the level of detail, we first compute each edge energy e_i by using equation 4.7. Each edge energy is normalise as follows.

$$\bar{e}_i = \frac{e_i}{a_i \times \beta_i} \tag{4.15}$$

where $\bar{e}_i$ indicates the normalised edge energy. a_i and β_i represent the height and width of the decomposed image at level i. The number of faces is counted by using a face detector described in [137].

4.4.2 Features for Level 2 Classification

A shot which is classified as a CDS or a MDS is going to another level of classification. In order to do that, different features are used for an effective result. In the following, features which are used in level 2 classification are described in detail.

Features for Close Distance Shot Classification

The CDS consists of three sub-classes of cinematographic shot - close-up shot, medium close-up shot and over-the-shoulder shot. For CDS classification, a 9 dimensional feature vector is used. In this case, we use three normalised horizontal context saliency features $\bar{\Omega}_i^h$. Again, we use horizontal regions with skin colour features S_i^h. We use the horizontal direction that gives a notion of constituent of the scene. For example, if the saliency values and the regions with skin colours are nearly evenly distributed in all three segments, the scene is more likely to be a close-up shot. We compute three more features from the edge energy E by segmenting the scene vertically into three equal sized regions. After computing the total local energies, these values are normalised using the total energy of E. The intention is to represent the spatial construction of a scene.

After extracting the features, a feature vector is created and used for SVM classification. The classification procedure has been described in Sub-section 3.5.2 of Chapter 3.

Features for Medium Distance Shot Classification

There are two subclasses of cinematographic shot belonging to medium shot: medium shot and cut shot. Medium shot and cut shot are classified by using a key feature - human face occurance. We adopt the method in [137] to identify faces in a keyframe. If there is a face appearing in a key frame then the shot is classified as medium shot, otherwise the shot is classified as a cut shot.

4.5 Experimental Results

4.5.1 Dataset Preparation and SVM Classification

In this chapter, we use the same dataset that we have prepared for Chapter 3. The dataset contains a total of 3206 keyframes representing six different classes, namely 1) Close-up (C) shot, 2)Over the shoulder (OTS) shot, 3) Medium close-up (MCU) shot, 4) Medium (M) shot, 5) Cut (C) shot, and 6) Long (L) shot. The keyframes are taken from a wide variety of movie genres and the ground truths are labelled manually. The same training and testing sets from Chapter 3 are used for SVM training and evaluation purposes. For classification, One-Against-All 3-fold cross validation SVM classification approach is used.

4.5.2 Evaluation

The performance of the proposed features are measured in this subsection. The classification performance is measured in terms of precision rates and recall rates and f_1 scores. Table 4.5.1 shows the precision and recall rates of level 1 shot clas-

Table 4.5.1: Recall and precision rates for level 1 classification

	CDS	MDS	LDS
Recall	89.23%	84.88%	90.77%
Precision	83.35%	77.06%	75.99%

Table 4.5.2: Confusion matrix for level 1 classification

	CDS	MDS	LDS
CDS	89.23%	4.06%	6.61%
MDS	11.99%	83.86%	3.15%
LDS	5.88%	3.24%	90.77%

sification. As it can be seen, the overall recall and precision rates are satisfactory. Table 4.5.2 shows the confusion matrix of level 1 classification. It shows that the confusion of the CDS is almost equally with the MDS and the LDS. However, for MDS, more shots are classified as CDS. Similarly, the confusion of the LDS is almost equally with the CDS and the MDS.

The recall and precision rates of level 2 classification are given in Table 4.5.4. The classification performance of OTS shot is very high with racall rate more than 92% and precision rate more than 80%. Again, the classification performance for CU shot is also high with recall rate around 78% and precision rate around 84%. The error is analysed using the confusion matrix shown in Table 4.5.4. The majority of confusion of the CU shots is with the MCU shots. Similarly, the majority of

Table 4.5.3: Recall and precision rates for level 2 classification.

	CU	OTS	MCS	MS	CS
Recall	78.01%	92.23%	88.90%	81.95%	78.19%
Precision	83.96%	80.35%	65.06%	85.01%	77.00%

Table 4.5.4: Confusion matrix for classification of six types of shots.

	CU	OTS	MCS	MS	CT	LS
CU	78.01	8.25	13.65	0.00	0.00	0.00
OTS	6.25	92.23	1.15	0.00	0.00	0.00
MCS	9.65	1.39	88.90	0.00	0.00	0.00
MS	0.00	0.00	0.00	88.00	11.75	0.00
CT	0.00	0.00	0.00	17.62	81.89	0.00
LS	0.00	0.00	0.00	0.00	0.00	90.77

confusion of the MCU shots are with the CU shots.

The f_1 score of CU, OTS, MCS, MS, CT and LS are 80.88, 71.05, 60.69, 67.41, 64.23 and 82.73 respectively. Table 4.5.5 shows the f_1-scores of six shot types. We compare the performance with the results reported in Chapter 3. Figure 4.6.1 shows the result comparison using histograms. Although the f_1-score does not improve for CU shots, f_1-scores for rest of the shot types has been improved significantly. The most significant accuracy is achieved for long shot with f-score 82.73. The accuracies for the rest are also significantly high. We further compare the result of the proposed framework with the classification results achieved by using proposed features without using hierarchical approach. Figure 4.6.2 shows the result comparison using histogram. Although the classification performance of CU

Table 4.5.5: F-measure for classification of six types of shots.

CU	OTS	MCU	MS	CT	LS
80.88	71.05	60.69	67.41	64.23	82.73

increase by a small margin, other classes are far bellow then the achieved results by the proposed method. In conclusion, the proposed hierarchical approach has significant influence in classifying cinematographic keyframes.

4.6 Discussion

As mentioned, there exist a number of approaches for cinematographic shot classification. However, our proposed approach is a novel effort in classifying cinematographic shots. As the other approaches and shot classes are different from our approach, the results can not be compared. However, in our previous work in [1] and in Chapter 3, we took a similar approach to classify cinematographic shot. In comparing to our previous work, using the method proposed in this chapter improves the classification performance significantly. The improvement is achieved because of different classification approach. Unlike our previous work, the classification is done in two broad steps. In the first step, shots are classified based on the distance of the camera from a subject. Due to this simplified but effective approach, shots can be classified into subclasses with more accuracy. As it is found in the result comparison, the classification for the long shots is improved greatly. The next level of classification only concentrates on classifying similar types of shots.

Due to simplification of the overall approach, most of the shots classification accuracy is improved significantly. Although, the overall accuracy has been improved, the processing time of the proposed method in this chapter is a concern. As a consequence of inclusion of more features, the hierarchical approach is slower than the method proposed in Chapter 3. For a 688×272 pixel keyframe, the feature extraction times are 1.79 and 1.22 second for the proposed method and the method proposed in Chapter 3 respectively. In the experiments we use personal a computer with Windows XP operating system, Intel Core i5 2.5 GHz, 4-GB memory. We use Microsoft Visual Studio 8.0 with OpenCV 2.3 for programming.

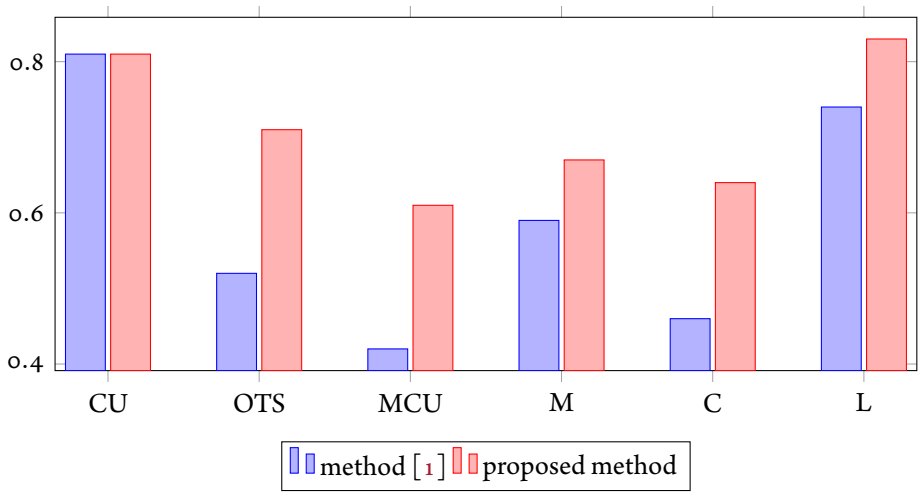


Figure 4.6.1: The comparison with our previous work proposed in [1].

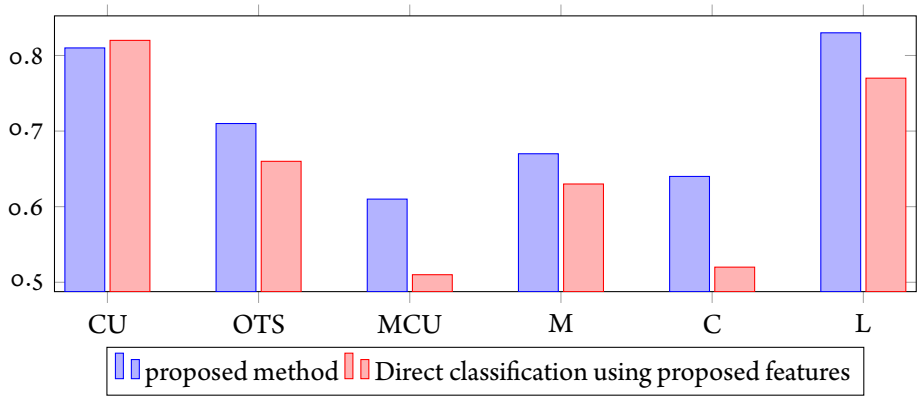


Figure 4.6.2: The classification performance of the proposed method and the method using the proposed features but without using hierarchical approach.

4.7 Summarisation

Due to the unorganised nature, the classification of cinematographic shot is not an easy task. In this paper, we propose a hierarchical cinematographic shot classification technique to classify six pre-defined classes. There are two steps in our classification technique. In the first step (level 1 classification) a shot is classified into one of the three broad classes. If the shot belongs to CDS or MDS class, the shot is further classified into detailed shot classes. CDS shots are further classified into CU shots, MCU shots and L shots. MDS shots are classified into M shots and C shots. For classification, we introduce a set of movie domain specific features. One of the important features is context saliency map. We also identify that the camera lens put signature on the shots depending on the distance of the objects from the camera lens. Thus, edge energies are computed and used as additional important features. Weighted hue histogram is computed to identify the dominant colour in the salient location of a keyframe. As the movie mostly involved

with human activities, the area with skin color is considered as another important feature. For feature extraction, we use photography composition rule (*i.e.*, rule of thirds) to represent the shot effectively. As there is no similar work done, we could not compare the results with other methods. However, compared with our previous work [1], the proposed method performs much better with significantly higher accuracies.

In the proposed method, we have not considered any motion information. Motion information provides an important clue in classifying shots. In the next chapter, we investigate the performance of shot classification including motion information.

Ever tried. Ever failed. No matter. Try again. Fail again. Fail better.

Samuel Beckett

5

Camera Motion Histogram Descriptor for Video Shot Classification

5.1 Introduction

In this Chapter, we propose a novel camera motion characterisation and description technique through identifying the camera motion patterns from the inherent motion structure buried in the raw video data. We concentrate on professionally captured videos (*e.g.,* feature films and sports) and thus, home videos are out of the scope of our work. We deal with motion information in two main stages, motion characterisation and motion description. Our intention is to exploit the redundant information of video data to characterise the camera motion patterns of video shots. Motion vectors (MVs) of consecutive frames are extracted and the

inconsistent motions are suppressed by applying statistical temporal motion analysis. Then, the global motion patterns of each video shot is represented by using a number of local motion descriptors. The local motion descriptors are used to describe the local camera motion patterns of different local regions. Then, the extracted features are used in a statistical learning framework to recognise the qualitative camera motion patterns.

At the motion characterisation stage, motion vector fields (MVFs) are constructed by extracting motion vectors from consecutive frames by using block matching (BM) technique. The MVFs are segmented into nine (3×3) equally sized local regions. Then, the temporal gradient of the motion vectors of each macro-block (MB) is computed. Then, by using an effective statistical measure on the gradient of motion, the motion vectors of interest (MVIs) are identified. The MVIs of each local region are then characterised by using principal component analysis. The most variance retaining principal component is identified to represent the camera motion compactly. To do this, we accumulate the MVIs from a small number of frames for a local region and construct a matrix. Then, the matrix is decomposed using SVD technique. Let us assume that we have t frames in a video shot. The matrix is formed by using MVIs of the corresponding local regions of n consecutive frames where $(n \ll t)$. Finally, the oriented angle of the principal component is computed and quantised with a predefined step size. The consecutive quantised angles of each local region are used to characterise the local temporal motion. At the motion description stage, quantised angles of each local region are used to create a histogram, where each local histogram is considered as local motion de-

scriptor. By combining all the local histograms, the camera motion histogram descriptor is formed for an input video. Figure 5.2.1 shows the flow diagram of the proposed method. The contributions of this chapter are summarised as follows.

- A novel compact camera motion characterisation technique is introduced to characterise the camera motion in a video.
- A novel shot level camera motion descriptor is proposed to represent the overall motion activity of a shot by using a histogram.

The detail of our contribution is described in Section 5.3.

5.2 Framework

Some approaches have been proposed to characterize camera motions. In [55], an approach was proposed to characterize camera motion in the temporal direction. A two dimensional affine motion model was applied on consecutive frames to estimate global dominant motion. Then, by assuming that the dominant motion happens due to camera motion, a qualitative description of dominant motion was estimated using the significance of the global affine components. Although the approach is useful for motion characterization in the temporal direction, it is not suitable for effective indexing purpose. A compact representation of motion is much desired for an effective solution. Ngo *et al.* [2] proposed an approach to classify video shots static, pan, tilt and zoom shot classes using spatio-temporal slice processing. The spatial dimension (x, y) at time t was processed by using spatio-temporal slices in the (x, t) space and (y, t) space. Then, the camera motions were

analyzed by utilizing Tensor histograms. As the authors assumed that there ware no object motions present around the boundary of the frames, the applications are only be limited to news and sports videos. Such assumption may cause poor performance in generalized video shot classification. Ewerth *et al.* [3] proposed another approach to classify video shots into basic camera motion classes by utilizing MPEG domain motion vectors. Firstly, motion vectors from only P-frames were extracted to represent the camera motions. Then, irrelevant motion vectors were identified and eliminated by applying two proposed filters, namely smoothness and neighborhood filters. Finally, camera motions ware modeled by using Nelder-Meade algorithm. Although their proposed method can be used to distinguish between translation and rotation around x and y axis, the success depends on the sample motion vectors and accuracy of outlier removal technique. In [146], a novel approach was proposed for camera motion analysis. In their proposed method, keypoints from consecutive frames were matched using Difference of Gaussians (DoG) and SIFT descriptors. Then, a voting process was applied to eliminate the foreground key points. Their method can be applied for video shot categorization. However, this method also lacks a compact representation of motions. A few optical flow based methods were proposed in [147–149]. Nguyen *et al.* [147] proposed a template based method to determine motion types from the computed optical flows. The applicability of template based methods to classify video shot is limited to slow videos. Almeida *et al.* proposed an optical flow based method to classify video shots into tilt, pan, roll and zoom classes. Their proposed method is considered to be a slow method. Moreover, the method lacks

a compact representation of the camera motions. In [150], a motion descriptor was proposed for motion activity for MPEG videos. Firstly, motion intensity were characterized into different intensity levels. Then, a histogram was computed using different intensity levels to describe a scene. The computed histogram was used as the scene descriptor. Although good performances were achieved in [150], direct use of MVs for compressed videos could be misleading. MVs in compressed domain may not represent the true optical flows.

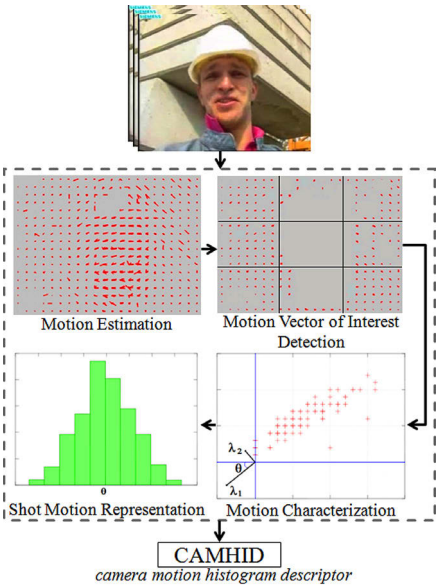


Figure 5.2.1: Flow diagram of the proposed camera motion histogram descriptor technique.

As far as video indexing and retrieval concern, both camera motion characterization and motion description can play an important role. Compact temporal motion characterization describes the camera motion pattern over time while a motion descriptor summarize the overall camera motion. Such characterization and

description techniques provide coherent motion information ranging from basic to complex camera movement. In this chapter, we proposed a camera motion descriptor for the purpose of motion based video indexing and retrieval. Figure 5.2.1 shows the flow diagram of the proposed method. In our proposed method, we include both motion characterization and motion description to address video indexing problem. In the proposed method, a compact camera motion characterization and description techniques are introduced by identifying the camera motion patterns from the inherent motion structure buried in the raw video data. We concentrate on the professionally captured videos (*e.g.*, feature films and sports) and thus, home videos are out of the scope of our work. We deal with motion information in two main stages, motion characterization and motion description. Our intention is to exploit the redundant information of video data to characterize the camera motion patterns of video shots. Motion vectors of each consecutive frames are extracted and the inconsistent motions are suppressed by applying statistical temporal motion analysis. Then, the global motion patterns of each video shot is represented by using a number of local motion descriptors. The local motion descriptors are used to describe the local camera motion patterns of different local regions. Then, the extracted features are used in a statistical learning framework to recognize the qualitative camera motion patterns. The proposed method works in four main steps: motion estimation, motion vector of interest detection, motion characterisation and shot motion representation. The detail of the proposed technique is described in the following subsection.

5.3 Proposed Camera Motion Descriptor

In this section, the proposed technique for camera motion descriptor is described in detail. We name the proposed descriptor the CAmera Motion HIstogram Descriptor (CAMHID). The CAMHID is constructed in four steps. Firstly, BM based ME technique is used to estimate the local motion of each MB. Then, by analysing the local motion of each MB in the temporal direction, MVIs are identified. Then, in the third step, MVIs are used to produce a sequence of compact representations of temporal motion. In the last step, the compact representations are used to obtain a normalised histogram as the local motion feature of video shots. As shown in Figure 5.2.1, the output of the above-mentioned four steps for feature extraction is a CAMHID, that integrates the motion features of the local regions. The following subsections describe the detail of CAMHID construction.

5.3.1 Block Matching Based Motion Estimation

BM based ME is a popular technique to estimate a local motion. This technique has been widely used for video compression, particularly for motion compensation in the current state-of-the-art video coding standards [151]. The ME techniques are used to find the optimal optical displacement in an MB of a frame. The optimal displacement is represented by a MV which corresponds to the coordinate displacements of the best matching block in the reference frame. For an MB in the i-th frame, MV is searched in the $(i+1)$-th frame. Let f^i and $f^{(i+1)}$ be two consecutive frames taken from a video shot. We construct a motion vector field

MVF^i by extracting all MVs belong to a frame. In order to do that, frame f^i is subdivided into non-overlapping MBs of size $N \times N$ (see Figure 5.3.1). For each MB, the most similar block in frame $f^{(i+1)}$ is identified by searching in the area of size $(M+N) \times (M+N)$ in $f^{(i+1)}$ as shown in Figure 5.3.1, where $M = 2 \times N$. For an MV $mv^i_{(x,y)}$, we need to compute the horizontal displacement $u^i_{(x,y)}$ and the vertical displacement $v^i_{(x,y)}$. Formally, we write:

$$(u^i_{(x,y)}, v^i_{(x,y)}) = arg \min_{u \in \{-\frac{M}{2}+1,\ldots,\frac{M}{2}\}, v \in \{-\frac{M}{2}+1,\ldots,\frac{M}{2}\}} e(x,y,u,v,i) \tag{5.1}$$

where,

$$e(x,y,u,v,i) = \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} |f^i(x+p, y+q) - f^{(i+1)}(x+u+p, y+v+q)|.$$

The optimal optical displacement identified by Eq. (5.1) is used to compute $u^i_{(x,y)} = (x-u)$ and $v^i_{(x,y)} = (y-v)$ which represent the horizontal and vertical displacements of the MV $mv^i_{(x,y)}$ respectively. Likewise, we construct $MVF^i = \{mv^i_{(x,y)}\}$, for $\forall x \in \alpha, \forall y \in \beta$. Here, α and β correspond to the width and height of the video frame respectively.

5.3.2 Detection of MV of Interest

In this subsection, we propose a technique to identify MVIs through analysing local motions in the temporal direction. Our goal is to identify the spatial region

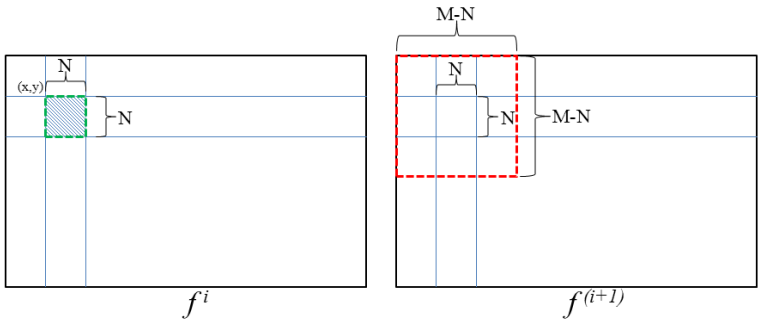


Figure 5.3.1: BM based MV searching from two consecutive frames f^i and $f^{(i+1)}$. For an $N \times N$ MB at (x, y) in f^i, the searching area is marked in $f^{(i+1)}$. The size of the searching area is $(M + N) \times (M + N)$ centring at the searching block region.

where the motion information has a direct relationship with the camera movement. The basic camera movements are categorised as static, pan, tilt, zoom and combination of them. The camera can be operated manually by holding in hand or by mounting on a tripod or any form of transportation. In professional video shooting (e.g., the shooting for sports, news, documentary and film), video shots are captured with smooth, jerking free and consistent camera motion. The objects in video frames can be static or dynamic. Due to camera movement, the MVs belonging to the static object region have a direct relationship with the camera movement. However, according to our observation, motion pertaining to non-rigid bodies and focused objects (e.g., objects are being shot) are independent of camera motion. Non-rigid bodies (e.g., rippling water and waving leaves) and players/actors often produce random and jerky motion with respect to camera frame. Figure 5.3.2 shows camera motion analysis by using benchmark video shots. Although the first 10 frames of the Foreman sequence is identified as with a static video shot, the subject actually produces a random motion with respect to the

camera frame. Other parts of the video mostly preserve the camera motion information. Similarly, the flower garden sequence mostly preserves camera motion information in the first 10 frames of the shot. The washed out areas of the helmet (in the Foreman sequence) and a big part of the sky area (in the flower garden sequence) do not produce any motion information due to inadequate detail of visual information. Based on this observation, in this subsection, we search for the MVIs where camera motion consistency is preserved in the temporal direction of the computed MVs. In order to do that, firstly for each MB, we compute the gradient of MVs in the temporal direction. The gradients of horizontal displacement $u^i_{(x,y)}$ and of vertical displacement $v^i_{(x,y)}$ are computed separately for each MB of the entire shot. Formally, we write:

$$\begin{aligned} \nabla u_{(x,y)} &= \{\nabla u^1_{(x,y)}, \nabla u^2_{(x,y)}, \ldots, \nabla u^{(t-1)}_{(x,y)}\}, \\ \nabla v_{(x,y)} &= \{\nabla v^1_{(x,y)}, \nabla v^2_{(x,y)}, \ldots, \nabla v^{(t-1)}_{(x,y)}\}. \end{aligned} \tag{5.2}$$

where,

$$\begin{aligned} \nabla u^i_{(x,y)} &= \left(u^{(i+1)}_{(x,y)} - u^i_{(x,y)}\right), \\ \nabla v^i_{(x,y)} &= \left(v^{(i+1)}_{(x,y)} - v^i_{(x,y)}\right). \end{aligned}$$

Next, MVs are determined as MVI or not by employing a simple and effective statistics based traditional measure of distance on the computed $\nabla u_{(x,y)}$, $\nabla v_{(x,y)}$ and computed motion vectors. We check the significance and consistency of the

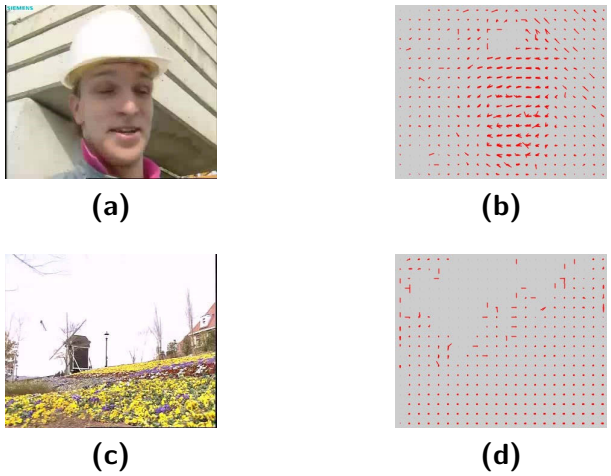


Figure 5.3.2: Motion analysis using benchmark video sequences. Subject and non-rigid body creates jerky motion with respect to the camera frame. (a) Foreman video sequence, (b) First 10 MVFs of the Foreman sequence are computed and superimposed, (c) Flower garden video sequence, (d) First 10 MVFs of the flower garden sequence computed and superimposed.

motion activity of MBs. The significance is checked by computing the mean magnitude of the MVs and consistency is checked by using the gradient of the MVs (described in Eq. (5.2)). If the MVs at (x, y) of frame f^t show that the motion is significant and consistent for k $\big(k < (t-1)\big)$ consecutive MBs in the temporal direction, then the MV is declared as an MVI. In this way, all MVIs are identified from the extracted MVs, provided that there is no MVIs extracted from the last

$k-1$ frames. The MVI determination process is formally written in Eq. (5.3).

$$mvi^{i}_{(x,y)} = \begin{cases} true & \text{if } \left(\mu_{dist^{i}_{(x,y)}} > \tau_1\right) \text{ and} \\ & \left(\left(\sqrt{\sigma^2_{\nabla u^{i}_{(x,y)}} + \sigma^2_{\nabla v^{i}_{(x,y)}}}\right) < \tau_2\right) \\ false & \text{otherwise} \end{cases} \tag{5.3}$$

where,

$$\begin{aligned} \mu_{dist^{i}_{(x,y)}} &= \frac{1}{k}\sum_{j=0}^{k-1}\sqrt{\left(u^{i+j}_{(x,y)}\right)^2 + \left(v^{i+j}_{(x,y)}\right)^2}, \\ \sigma_{\nabla u^{i}_{(x,y)}} &= \sqrt{\frac{1}{k-1}\sum_{j=0}^{k-1}\left(u^{i+j}_{(x,y)} - \mu_{\nabla u^{i}_{(x,y)}}\right)^2}, \\ \sigma_{\nabla v^{i}_{(x,y)}} &= \sqrt{\frac{1}{k-1}\sum_{j=0}^{k-1}\left(v^{i+j}_{(x,y)} - \mu_{\nabla v^{i}_{(x,y)}}\right)^2}, \end{aligned}$$

$\mu_{\nabla u^{i}_{(x,y)}} = \frac{1}{k}\sum_{j=0}^{k-1}\nabla u^{i+j}_{(x,y)}$, $\mu_{\nabla v^{i}_{(x,y)}} = \frac{1}{k}\sum_{j=0}^{k-1}\nabla v^{i+j}_{(x,y)}$, τ_1 is a threshold which is close to 0 and τ_2 is a threshold for motion inconsistency tolerance. Both thresholds are set experimentally (detail of the experiment is described in subsection 5.4.1). Figure 5.3.3 illustrates the MVI determination procedure from different camera motion types.

Figure 5.3.3: Illustration of MVI determination from different camera motion types. First row of each sub-figure represents 1st, 5th and 10th frame respectively of the video sub-shots. The left image of the second row of each sub-figure represents accumulated motion vectors from 10 consecutive frames of the corresponding sub-shots. The right image of the second row of each sub-figure represents the location of the MVIs determined by applying Equation (5.3).

5.3.3 Motion Characterisation

The MVI based motion characterisation technique is described in this subsection. Our intention is to characterise a video shot in such a way that preserves the motion content in a compact manner. Figure 5.3.4 shows roughly a region-wise shot motion summary of common video shot types (according to the definition of each shot type). It shows that the motion patterns of static, tilt and pan shot are similar in every local region. However, zoom shot's motion pattern varies and is local region dependent. Therefore, in order to identify the characteristics of a video shot, we need to consider this fact. Accordingly, we divide the computed MVFs into nine (i.e., 3×3) non-overlapping local regions of equal size. The MVF in region (p, q) of frame i is denoted by $MVF^i_{(p,q)}$, $p \in \{1, 2, 3\}$ and $q \in \{1, 2, 3\}$. For each region, $MVF_{(p,q)}$'s motion contents in the temporal direction are separately and compactly represented. At the end, the compact representation of camera motions of all local regions are combined to characterise the whole shot's motion. Figure 5.3.5 shows the basic strategy of the motion characterisation procedure. As shown in the figure, $mvi^i_{(x,y)} \in MVF^i_{(p,q)}$ of n consecutive temporal regions are accumulated for compact representation. During the accumulation, we count that if we have a significantly big number of MVIs for characterisation. If the number is lower than 10% of the total number of MVs of a local region, then we conclude that this happens due to lack of enough camera movements and/or random motion from the non-rigid body. Hence, the camera motion related to this particular n $MVF^i_{(p,q)}$ is identified as static. Otherwise, the motion related to the region is

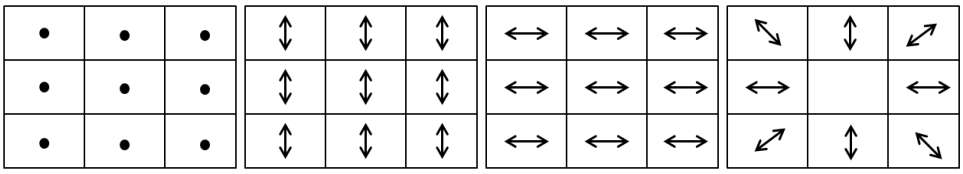


Figure 5.3.4: Region-wise shot motion summary of static, tilt, pan and zoom shot (left to right). Each of the local region represents the rough direction of a camera motion.

characterised and compactly represented.

The compact representation is accomplished using the singular value decomposition (SVD) technique [152]. At the beginning, a matrix A is created with the MVIs from n local regions of $MVF^{i}_{(p,q)}$ accumulated using raster scan method. Let matrix A contain l MVs and the MVs are d dimensional vectors. In our case $d = 2$ as we have only two components in the motion vectors. Therefore, A is a $l \times 2$ matrix. Then, we apply SVD on A to decompose it as follows.

$$A = U\Lambda V^{T} \tag{5.4}$$

where U is a $(l \times 2)$ orthonormal matrix. The columns of U are the eigenvectors of AA^{T}. $\Lambda = diag(\lambda_1, \ldots, \lambda_d)$ is a $d \times d$ diagonal matrix containing the singular values in descending order. The singular values are the square roots of the eigenvalues of both AA^{T} and $A^{T}A$. The magnitude of each singular value corresponds to the importance of the corresponding principal component. V is a $(d \times d)$ orthonormal matrix and the columns of V are the eigenvectors of $A^{T}A$. We are interested in V as it encodes the coefficients used to expand A in terms of U. As the top s $(s < d)$ principal components approximate a significant amount of information

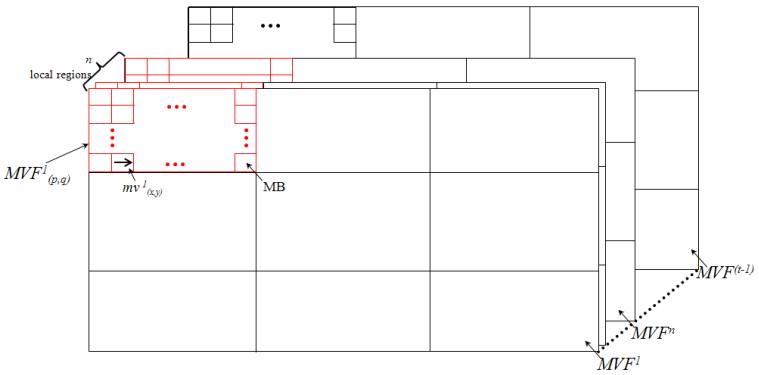


Figure 5.3.5: Local motion characterisation. The MVIs of n consecutive regions are accumulated for a compact representation.

of the original data [153], we represent the camera motion using the most dominant principal component of V. Therefore, the accumulated MVI from n blocks of $MVF^{i}_{(p,q)}$ is compactly represented by this most dominant component. The vector is identified as $pc^{(i,n)}_{(p,q)}$, where n is a constant.

5.3.4 Shot Motion Representation

In the previous subsection, each motion is either characterised as static or further characterised according to the direction of the most dominant principal component. For each local region, the results obtained in the previous step shown in Subsection 5.3.3 to form a histogram. First of all, the oriented angles of the principal components $pc^{(i,n)}_{(p,q)}$ are computed. Each angle is computed with respect to x

axis. Eq. (5.5) is used to compute the oriented angle.

$$\theta_{(p,q)}^{(i,n)} = \begin{cases} \cos^{-1}\left(v \cdot pc_{(p,q)}^{(i,n)}\right) & \text{if } y_{pc} \geq 0 \\ 360^\circ - \cos^{-1}\left(v \cdot pc_{(p,q)}^{(i,n)}\right) & \text{otherwise} \end{cases} \tag{5.5}$$

where y_{pc} is the y component of $pc_{(p,q)}^{(i,n)}$ and v is an unit vector along x axis. The SVD technique is susceptible to noisy data. However, a rough estimation of the motion angle is sufficient enough for our task. The computed angle is quantised with Q levels in the range $[0^\circ, 360^\circ]$. Figure 5.3.6 shows the angle quantisation strategy. The first angle level range is 345° to 15° in counter clockwise direction, and rest of the levels are equally spaced along the angle circle. Using the computed angle and the static motion information, The histogram for a region is formulated as follows.

$$H_{(p,q)}(c) = \left(h_{(p,q)}^0(c), h_{(p,q)}^1(c), \ldots, h_{(p,q)}^Q(c)\right) \tag{5.6}$$

where $h_{(p,q)}^i(c)$ represents the count (or height) of the i-th bin. The first bin represents static region count. The rest are for the quantised angle index $i \in \{1, \ldots, Q\}$. Finally, the histogram is normalised for a uniform representation using the following equation.

$$\hat{H}_{(p,q)}(c) = \left(\frac{h_{(p,q)}^0(c)}{C}, \frac{h_{(p,q)}^1(c)}{C}, \ldots, \frac{h_{(p,q)}^Q(c)}{C}\right) \tag{5.7}$$

where $C = \sum_{i=0}^{Q} h_{(p,q)}^i(c)$. Eq. (5.7) is used as the feature of a region to represent the camera motion related to the region.

After all mentioned in Subsections 5.3.1-5.3.4, we integrate all of the local his-

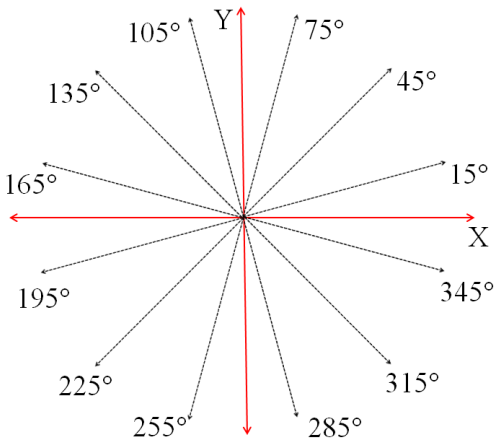


Figure 5.3.6: Illustration of compactly represented motion quantisation rule. The angle of each $pc_{(p,q)}^{(i,n)}$ resides in one of the 12 ranges.

togram features to form a single feature vector. This vector representation is considered as the camera motion histogram descriptor for an input video shot.

5.4 Experimental Results

The performance of the proposed camera motion histogram descriptor is experimented and evaluated in this section. The proposed descriptors are extracted from the training datasets to train the support vector machine (SVM) classifiers. Then, the performance is evaluated based on the classification results on the testing datasets. The following subsections describe the detail of the datasets preparation and the evaluation procedure.

5.4.1 Dataset Preparation and Feature Extraction

As discussed in Chapter 1, based on camera motion characteristics, video shots are mainly classified into four basic classes: 1) static, 2) pan, 3) tilt, and 4) zoom. The static shots are captured by placing or holding the camera firmly without any significant camera movement. For the close-up view of objects, static video shots are often captured. The panning shots are captured by rotating the camera about the vertical axis. For the purpose of following objects horizontally, pan shots are often used. Tilt shots are captured by rotating the camera about the horizontal axis. Tilt shots are used for following objects in the vertical direction. Zoom shots are captured by changing the focal length of camera lens. In order to show close-up view from the long view of the object or vice versa, zoom shots are often used. The shots are labelled based on the observation defined above. A video shot is captured continuously, and it may contain different types of camera motions. In that case, the label is given based on the most dominant camera motion. A shot without any dominant camera motion is labelled as *others* shot.

To evaluate the classification performance, we conduct experiments on two datasets that we have created. The first dataset (Dataset 1) is created from cinematographic video shots. Dataset 1 consists of training dataset and testing dataset. The training dataset size is selected by analysing the classification performance on different sizes of data. Figure 5.4.1 shows the learning curves on the training data and the cross validation data. As it can be seen, with the increasing size of the training data, mean error rate is decreasing and eventually comes to an optimal state. From this analysis, we select the optimum size of training data size from each class

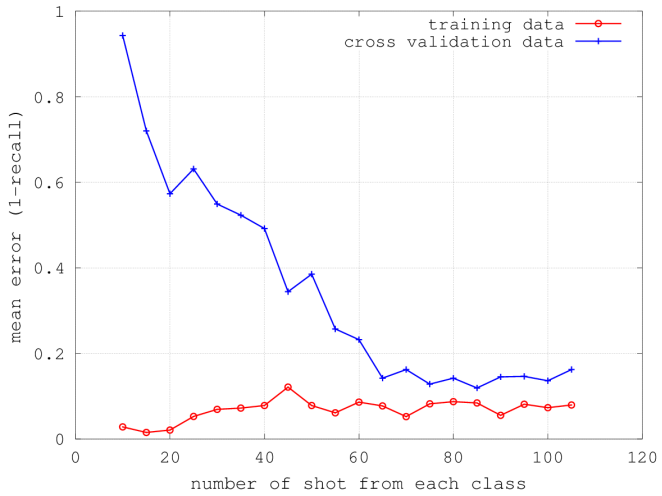


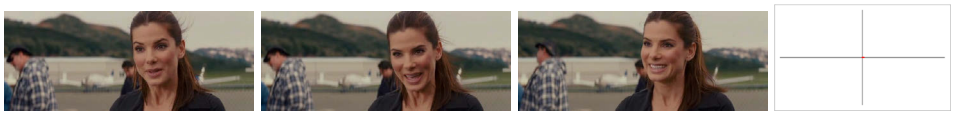
Figure 5.4.1: Training dataset size selection by analysing the classification performance on different dataset size from each shot class. For this experiment $k = 10$ and $n = 5$ are used.

to train the SVM classifiers. There are 425 shots finally taken (85 shots from each class) from three Hollywood films (The Proposal, Mission Impossible II and The Terminal). In order to create testing data, the shots are taken from three other Hollywood films (The Lord of The Rings I, A Beautiful Mind and Hotel Rwanda). The detailed breakdown of the testing data in Dataset 1 is given in Table 5.4.1. Figures 5.4.2 (a) - (e) show examples of different shot types in Dataset 1.

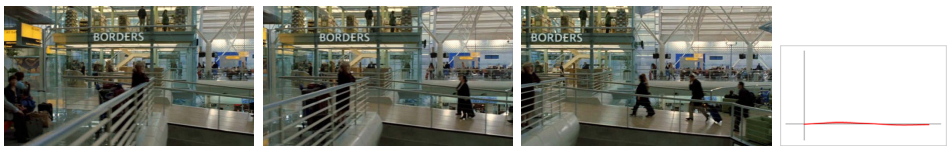
Table 5.4.1: Detailed breakdown of the testing data in Dataset 1. Dataset 1 is collected from Hollywood films.

	static	pan	tilt	zoom	other
no. of shots	1623	229	97	88	396

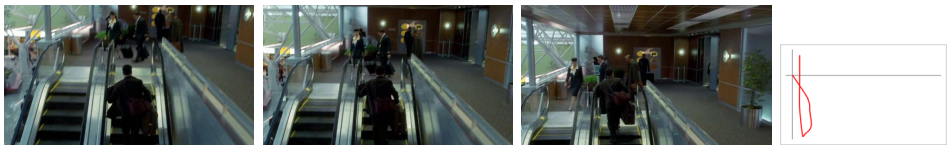
Another dataset (Dataset 2) is created from soccer game video. In sports videos,



(a) static shot



(b) pan shot



(c) tilt shot

(d) zoom shot

(e) other shot

Figure 5.4.2: Examples of video shot classes in Dataset 1. Each sub-figure contains the 1st, one of the intermediary and the last frames. The right image of each sub-figure represents the overall motion trajectory (the origin is the starting point). Zoom shot contains four motion trajectory as using one figure it is not possible to represent overall motion trajectory of a zoom shot.

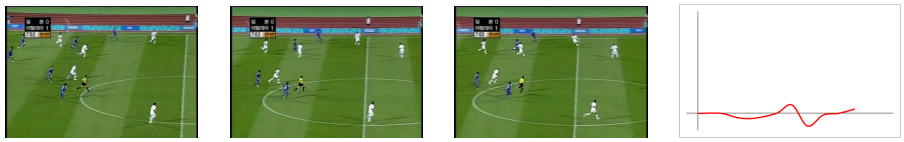
particularly in soccer videos, panning directions have an important clue of indexing. Panning directions are used to identify the attacking team in an attack shot. Therefore, panning preserves some degree of semantics in sports videos. Based on the panning directions, pan shots can be sub-divided into pan-left and pan-right subclasses. Video shots which cannot be classified according to the definition are tagged as *other* shots. Similar to Dataset 1 preparation, we create a training set by analysing the learning curves. We form the training dataset consisting of 150 shots (50 shots from each class) from two soccer videos (Korea vs. Germany and Japan vs. Italy). Then, for testing the classification performance, a testing dataset is created. The shots are taken form three soccer videos (Manchester United vs. Tottenham, Chelsea vs. Liverpool and Korea vs. Kuwait). The detailed breakdown of the testing data in Dataset 2 is given in Table 5.4.2. Figures 5.4.3 (a)-(c) show examples of different shot types in Dataset 2. Both datasets are manually segmented and labelled.

Table 5.4.2: Detailed breakdown of the testing data in Dataset 2. Dataset 2 is collected from soccer game videos.

	pan left	pan right	other
no. of shots	824	751	398

While extracting features from a video shot, we set the MB size to be of 16 $\times$ 16 pixels. The size of k and n is crucial for overall performance of the proposed method. Smaller size of k unstable overall MVI determination procedure because of noisy/random motion. On the other hand, a bigger size of k includes multi-

(a) left pan shot



(b) right pan shot

(c) other shot

Figure 5.4.3: Examples of video shot classes in Dataset 2. Each of the sub-figure contains the 1st, one of the intermediary and the last frames. The right image of each sub-figure represents the overall motion trajectory (the origin is the starting point).

ple camera motion in the MVI determination procedure which makes it difficult to identify the true MVI. Similarly, size of n also affect the overall performance of the proposed method. In order to determine the optimum size of k and n, the values are set experimentally based on the classification accuracy rate on Dataset 1. The accuracy rates are measured by using different k and n values and the best performing values are set for the rest of the experiment. Figure 5.4.4 shows the classification accuracy rates based on experiment. According to the performance, we set $k = 10$, and $n = 5$. The consistency and significance threshold selection is also important to achieve the best performance of the proposed method. The thresholds τ_1 and τ_2 are also set experimentally. Figure 5.4.5 shows the classifica-

tion accuracy rates based on the experiments conducted on Dataset 1 for changing thresholds settings. As it can be seen, for $\tau_1 = 1.5$ and $\tau_2 = 0.35$, we have achieved the optimum results. In the proposed camera motion histogram descriptor, the local region features mainly describes the camera motion of the corresponding local region. To determine the effectiveness of different local regions' motion descriptor, a feature selection experiment is conducted. Figure 5.4.6 shows f_1-scores on different feature set extracted from Dataset 1. As it can be seen, with the changing size of the training data size, the f_1-scores using four corner regions outperforms the rest. Accordingly, instead of considering all the local features, we only consider the features from the prime corner regions (top left, top right, bottom left and bottom right) in rest of the experiments. From all four corner regions, we compute their features and put them sequentially to obtain the shot-camera motion descriptors.

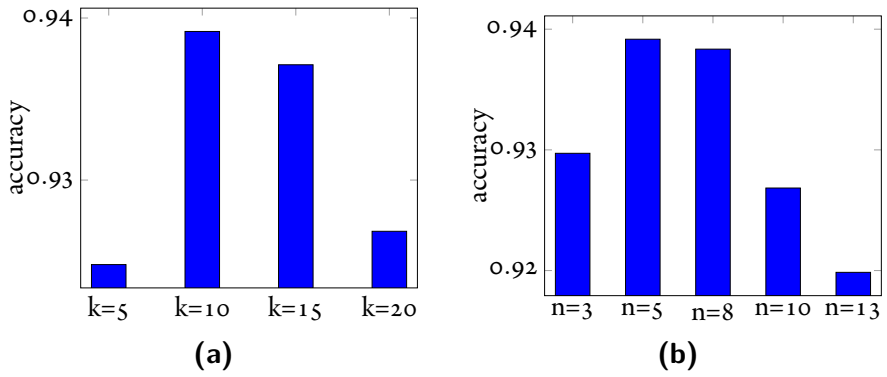


Figure 5.4.4: Classification accuracy measurement on Dataset 1. (a) Average classification accuracy for a changing k values and keeping $n = 3, 5$ and 8 and (b) average classification accuracy for a changing n values and keeping $k = 5, 10$ and 15. For both of the case τ_1 and τ_2 are set to 1.0 and 0.5 respectively.

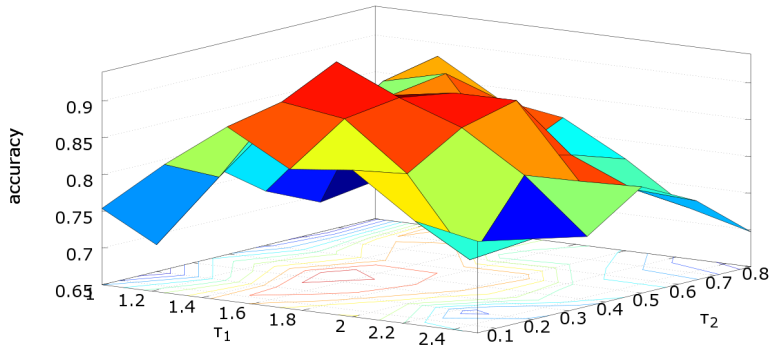


Figure 5.4.5: Classification performance using different threshold values in Equation (5.3). As it can be seen, the best classification accuracy is achieved by using $\tau_1 = 1.5$ and $\tau_2 = 0.35$.

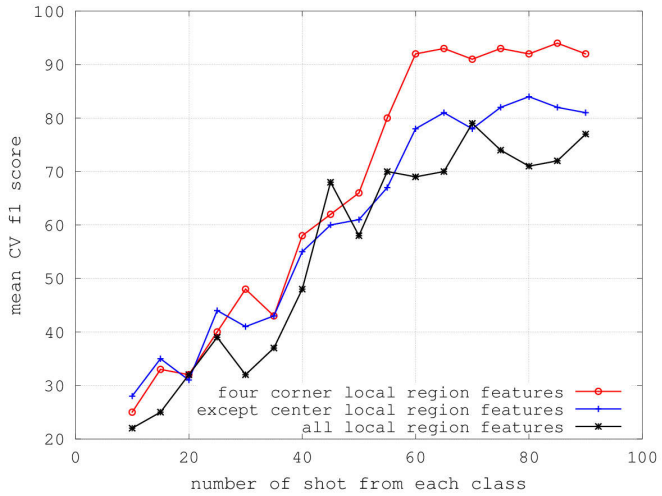


Figure 5.4.6: Shot classification performance based on different feature sets.

5.4.2 SVM Classification

In this chapter, we adopt One-Against-One approach for video shot classification. The effectiveness of One-Against-One approach has been presented in [142]. To summarise the One-Against-One multi-class SVM, let us assume that we have n training data in d dimensional space belonging to m classes $\{\mathbf{x}^i, y^i\}, \quad \mathbf{x}^i \in \mathbf{R}^d, \quad i = 1, \ldots, n, \quad y^i \in \{1, \ldots, m\}$. This approach constructs $m(m-1)/2$ classifiers using the training dataset. Each of the classifier is obtained by using the training data of the corresponding two classes. For class i and class j, the binary classification problem is formally written as:

$$\begin{aligned} &\min_{w_{ij}, b_{ij}, \xi_{ij}} \quad \left(\frac{1}{2}(w_{ij})^T w_{ij} + C\sum_t \xi_{ij}^t (w_{ij})^T\right) \\ &(w_{ij})^T \varphi(x^t) + b_{ij} \geq 1 - \xi_{ij}^t (w_{ij})^T, \text{if } y^t = i \\ &(w_{ij})^T \varphi(x^t) + b_{ij} \leq -1 + \xi_{ij}^t (w_{ij})^T, \text{if } y^t = j \end{aligned} \tag{5.8}$$

where ξ_{ij}^t is a non-negative slack variable, $\varphi(x^i)$ is a function to map x^i into a higher dimensional space and C is the penalty parameter. By minimising $\frac{1}{2}(w_{ij})^T w_{ij}$, we want to maximise the margin, $\frac{2}{||w_{ij}||}$, between class i and class j. The penalty term $C\sum_t \xi_{ij}^t$ is used to reduce the number of training errors for linearly non-separable cases. The goal is to find an optimal separating hyperplane by obtaining a balance between the regularisation term $\frac{2}{||w_{ij}||}$ and the training errors. To improve the separability, the data are mapped into a higher dimensional dot product space using a kernel function $K(x^i, x^j)$. One of the such functions is radial basis function (RBF)

and is used in this work. The RBF kernel is expressed as

$$K(x^i, x^j) = e^{-\gamma||x^i, x^j||^2} \tag{5.9}$$

where γ is the width control parameter. The accuracy of the SVM classification depends on the values of two parameters C and γ. Careful selection of these two parameters is important otherwise the classifier may perform poorly in the testing phase. A cross-validation approach is commonly used to determine the best parameters. We find the best penalty parameter C from the range $\{2^{-5}, 2^{-4}, \ldots, 2^4\}$ and width control parameter γ from the range $\{2^{-2}, 2^{-1}, \ldots, 2^7\}$.

Once the training is accomplished, the testing is done using the voting strategy called "Max Wins", proposed in [154]. In summary, for each comparison given data x, the sign of $((w_{ij})^T \varphi x + b_{ij})$ indicates the class of belonging. If the sign indicates that x belongs to class i, then the vote of class i is increased. Otherwise, the vote of class j is increased. At the end, the class with maximum vote is declared as the class of x. In case of draw, lowest index class is considered as the winner.

5.4.3 Evaluation

The performance of the proposed descriptor is evaluated by using SVM classifier. For both datasets, the effectiveness is shown by using confusion matrix and by computing recall, precision and f_1-score rates. Although there are wide variety of shot classification methods proposed in the literature, we choose [2] and [3] to be compared with our results with due to the direct relevance of the classified classes.

In Dataset 1, cinematographic shots are classified into 5 classes: static, pan, tilt,

zoom and other. The SVM classifier is trained using the training dataset of Dataset 1 and the performance is evaluated using the testing dataset. Table 5.4.3 shows the confusion matrix of the shot classification performance on Dataset 1. The recall, precision and f_1-score rates are reported in Table 5.4.4. As shown in the table, the correctly classification performance is reasonably high. We compare the performance of shot classification with two existing approaches. In [2], video shots are classified into static, pan, tilt and zoom classes. Although the achieved classification accuracy is very high, their own created dataset size is very small (consisting of 45 shots only). Another approach is reported in [3] where video shots are classified into three classes (pan, tilt and zoom) based on camera movements. In this case, the authors used a dataset of only 32 MPEG-1 video sequences, which is also considered as a very small data set. In order to prove the effectiveness of our approach, we have created Dataset 1 with reasonably and significantly bigger number of shots. Shot classification results are compared with [2] and [3]. Figure 5.4.7 shows the recall, precision and f_1-score comparison. The average recall and precision rates of [2] are 89.29% and 88.0% respectively and rates of [3] are 97.66% and 90.03% respectively. As the datasets used in [2] and [3] are not made available for public access, we cannot directly compare our results with the results using the two methods. However, our results demonstrated above have been promising and consistent. The average recall and precision rates of the proposed method are 93.6% and 89.8% respectively. Moreover, we also compare the f_1-scores which shows consistency of the classification performance. The average f_1 scores of [2] and [3] and the proposed method are 88.08%, 93.62% and 91.52% respectively.

Although the proposed method could not outperform the other two methods, the performance of the proposed method is considered as more acceptable because of the higher complexity level of our datasets. The difficulty level of [2] and [3] are considered at a minimal level. The data used in [2], consists of indoor shots only. In [3], no specific description was given about the dataset (except they used 32 MPEG video sequence). Compare to their datasets, our datasets are richer in motion content. The shots consist of wide range of shooting sets with complex camera motion. The cinematographic shots are often captured with the help of advanced tools (e.g., truck, crane and dolly) and technologies. Such things help the directors to enjoy more freedom to capture shots than sports, news and home videos. As a consequence, the cinematographic shots contains more complex camera motion patterns. In a summary, our created Dataset1 is not only large datasets, but also contain wide range of complex camera motion. Considering this fact, the performance of our approach is considered as consistent can be applied in any professionally captured video indexing purpose.

To justify our claim, the performance of shot classification is further evaluated using Dataset 2 to show the consistent performance of our approach in another type of videos - sports videos. After training the SVM Classifier using the training data of Dataset 2, the performance is evaluated using the testing data of Dataset 2. The confusion matrix is reported in Table 6.3.4 and recall, precision and f_1-score performance are given in Table 5.4.6. As shown in these two tables, the classification performance is again promising on this very different dataset from Dataset 1. According to the experiment, it is clear the capability of the proposed descriptor.

It can perform effective classification on any camera motion dataset by training the from the desired classes.

Table 5.4.3: Confusion matrix of shot classification using Dataset 1. Values within the parenthesis indicate the number of shots.

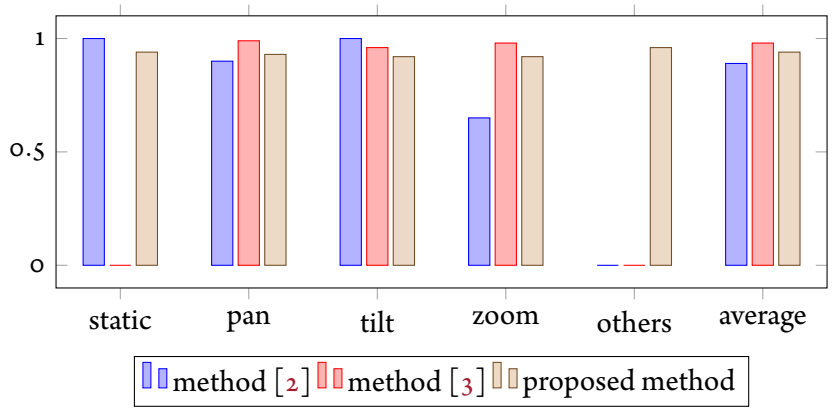
	static	pan	tilt	zoom	other
static (1623)	0.94 (1520)	0.01 (19)	0.00 (8)	0.00 (6)	0.04 (70)
pan (229)	0.00 (1)	0.93 (213)	0.00 (1)	0.00 (0)	0.06 (14)
tilt (97)	0.00 (0)	0.00 (0)	0.92 (89)	0.01 (1)	0.07 (7)
zoom (88)	0.01 (1)	0.01 (1)	0.00 (0)	0.93 (82)	0.05 (4)
other (396)	0.03 (10)	0.00 (1)	0.01 (3)	0.00 (1)	0.96 (381)

Table 5.4.4: Recall, precision and f_1-score measures of shot classification performance using Dataset 1.

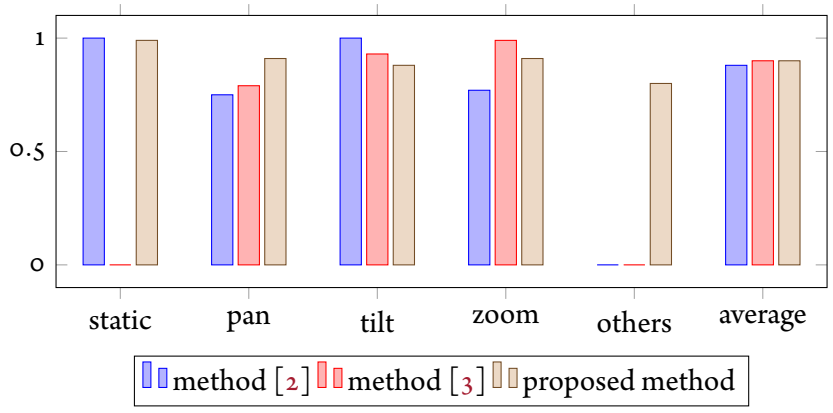
	static	pan	tilt	zoom	other
recall	0.94	0.93	0.92	0.93	0.96
precision	0.99	0.91	0.88	0.91	0.80
f_1-score	0.96	0.92	0.90	0.92	0.87

Table 5.4.5: Confusion matrix of shot classification using Dataset 2. Values within the parenthesis indicate the number of shots.

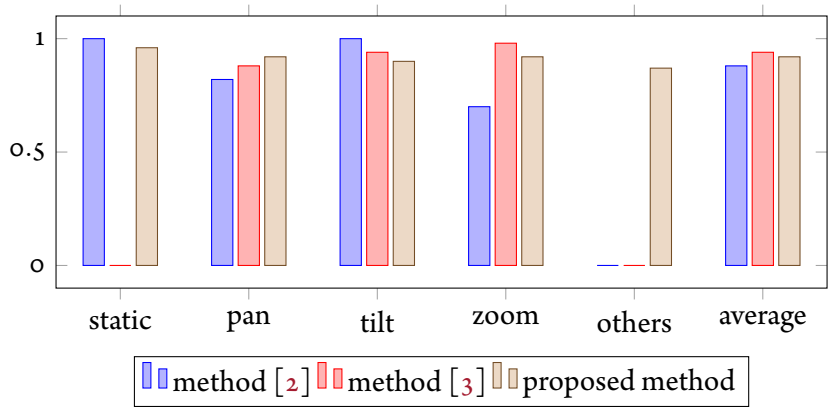
	pan left	pan right	other
pan left (824)	0.97 (798)	0.00 (4)	0.03 (22)
pan right (751)	0.00 (2)	0.96 (720)	0.04 (29)
other (398)	0.05 (18)	0.06 (24)	0.89 (356)



(a) recall



(b) precision



(c) f_1-score

Figure 5.4.7: Video shot classification result comparison with [2], [3]. (a) Recall rate comparison, (b) Precision rate comparison, (c) f_1-score comparison.

Table 5.4.6: Recall, precision and f_1-score measures of shot classification performance using Dataset 2.

	pan left	pan right	other
recall	0.97	0.96	0.89
precision	0.98	0.96	0.87
f_1-score	0.97	0.96	0.88

5.5 Summarisation

In this chapter, a novel non-parametric camera motion descriptor has been proposed. Using the proposed method, videos are classified according to the basic qualitative camera motion patterns. In order to do that, firstly, the temporal qualitative camera motion has been characterised using a series of compactly represented vectors. Then, the local camera motion of a video shot is described using a number of histograms. Finally, by combining the local histograms, overall camera motion patten of a video shot is described. One main advantage of the proposed descriptor is its versatility. Any particular camera motion type can be detected by training the classifier to identify that particular camera movement. We have evaluated the performance of the proposed descriptor and compared our approach with two existing approaches. It has been demonstrated that the proposed descriptor has a strong capability to effectively discriminate different types of camera movements on different types of videos. The proposed approach performs robustly on the video shot where the background is visible. However, in a extremely close-up video shot case, most of the background is hidden by an object and the proposed method fails to classify the shot to the true camera motion class.

The whole problem with the world is that fools and fanatics are always so certain of themselves, and wiser people so full of doubts.

Bertrand Russell

6

Application of CAMHID in Cinematographic Shot Classification

6.1 Introduction

In this Chapter, we extend CAMHID to perform the cinematographic shot classification task, which involves classifying cinematographic shots into the film directing semantic classes. In order to do that, we extend the feature space by extracting more features which considers the depth of a shot. In the following, we first describe the cinematographic shot framework and describe the directing semantic classes. We also discuss the need of additional features which enhance the discriminating capability in cinematographic directing semantic classes.

The contribution of this Chapter is as follows.

- We extend the CAMHID features using the existing MVI to incorporate the rough depth information.
- We investigate the performance of proposed camera motion descriptor along with a set of additional features in cinematographic shot classification.

6.2 Features Extraction for Cinematographic Shot Classification

Film making is completely based on the film making grammars. The directors heavily apply these film making grammars on every single cinematographic shots. The directors' main intention is to visualise the screenplay by capturing a cinematographic shot through a set of camera motions and a set of viewpoints. Capturing grammatically correct cinematographic shots ensures the viewer attentions on the predetermined actor(s), object(s) or place(s) based on the screenplay (henceforth, we will use 'object' and 'actor' interchangeably). According to [4], two of the major issues which determine the viewer attentions are as follows.

- Camera operation: a set of well defined camera operations are routinely performed to ensure the presence of different actions from the object of interest on the view plane. The camera operation is a strong indication of the categories of happening from the directing point of view. For example, static shots are often used to display the emotion of the actors or panning shots are often used to make sure the presence of object of interests on a view plane.

- Camera distance: The size of the objects of interest on a view plane carries different semantics from the direction sense. In cinematography, wide shots are often used to relate the object of interest with the surrounding environments while close-up shots are often used to display the emotional aspects on the actor's faces.

Based on these two issues, the construction of a taxonomy of the cinematographic directing semantics is discussed in the following. In directing semantic classes, the quality of camera operation and camera distance are more important than their quantity. For example, the differentiation between a slow zooming and a fast zooming is a subjective matter and quantitative measurements can be another research topic. Hence, in this chapter, we only consider the qualitative camera motion and distance. The relationship of the camera motion and object distance is important in directing semantic classes. The presence of a focused object makes the viewers feel like they are tracking the object. For example, a panning shot with a focused object gives a feeling to the viewers that the viewers personally track the object. However, without any focused object, a panning shot simply introduces a place to the viewer. In cinematography, this type of shots is only used to establish a new setting influencing viewers' mind. Scene composition is another aspect of cinematography which handles different issues such as distance of camera, camera angle and light. Among them, distance of camera is crucial as it determines the degree of emotional involvement of a viewer. In movies, we often see that highly emotional scenes are presented by using close-up/medium shots. Long distance shots are used to establish the context of a focused object. For the task of cinemato-

graphic shot classification, we group close-up and medium shots into one class as it is not easy to distinguish the purposes of using these two types of shots. In a wide range of movies, the use of close-up and medium shots are very similar for similar emotional shots. However, long shots are mainly used for contextual tracking and contextual establishments.

6.2.1 Cinematographic Directing Semantic Classes

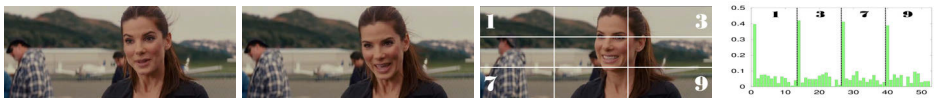
In Chapter 1, cinematographic directing semantic classes are classified which include stationary, contextual-tracking, focus-tracking, focus-in, focus-out, establishment, and chaotic shots. Stationary shot (S) comprises a significant portion of cinematographic shots. This type of shots is mainly used in dialogue shots. This type of shots contains minimum amount of camera movements to concentrate the viewer attentions on the actor's activities. Figure 6.2.1(a) shows an example of static shots and the corresponding CAMHID features. In this particular case, as it can be seen, the shot is captured by focusing on the actress using close-up shot while the camera movement remains almost static. Tracking shots are the one which are captured by focusing on object(s) and follow along the direction of the movement of the object(s). This type of shots is used to closely relate viewers to the objects [21]. It makes viewers feel like they are following the objects. Because of its own characteristics, tracking shots are considered as an important shot class. There are two types of tracking shots used in cinematography. Contextual tracking (CT) shots which establish a relationship of an object with the context by capturing a bigger picture of the scene. The focused object is captured using a

long shot so that the object looks smaller but provides scenic detail of the shooting set by using panning camera movements. Figure 6.2.1(b) shows an example of contextual tracking shot along with its CAMHID feature where the actress is being shot with a clear indication of the context (cityscape view). Focus tracking (FT) is another variant of tracking shots which provides a closer view of the objects. The intention behind taking this type of shots is to focus on the closer detail of the object while tracking. Figure 6.2.1(c) shows an example of focus tracking shots along with its CAMHID features. In cinematography, focus-in (FI) shots are captured in two ways. Firstly, zooming in by shortening the focal length of the camera lens and secondly, moving the camera to the object to shorten the camera distance for a closer view of the object. Both of these are mainly used to provide a greater detail of a focused object to highlight some important detail. Figure 6.2.1(d) shows an example of focus-in shot, where the object is getting bigger by changing the focal length. Focus-out (FO) shots are used to detach emotional involvement of the viewers from an object or relax the viewers by changing the viewing space. This effect is usually achieved through zooming out or dolly out shots, as the camera gradually moves away from the subject and creates emotional distance. Figure 6.2.1(e) shows an example of focus-out shot and its CAMHID features. The shot was captured by changing the position of camera distance.Establishment shots (ES) form another important directing semantic class which is used in cinematography. This type of shots is used to introduce a location to establish a relationship with the following sequence of shots. This type of shots is often taken by panning the camera without focusing on any particular object. Figure 6.2.1(f) shows an example of

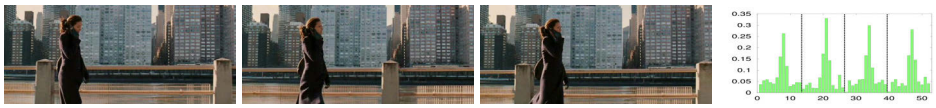
establishment shots and the corresponding CAMHID features. Chaotic shot(C) are characterised by the chaotic movement of the camera to follow an object or an object action. Chaotic shots are the ones which cannot be characterised as anyone of the above mentioned classes. Generally, a random camera motion happens due to focusing on an object's random motion. In order to represent fast action (or motions), directors apply this technique. In this shot type, it is not usually for the fast moving object to dominate viewer attention. Such shots are usually used to represent thrills and used more often in action films. Figure 6.2.1(g) shows an example of chaotic shots and its CAMHID features.

6.2.2 Feature Extraction from Cinematographic Shot Classes

The far right column of Figure 6.2.1 shows the CAMHID features of each camera motion type. Figure 6.2.1 (a) shows an example static shot and the corresponding CAMHID which combines the features from the four prime corner regions. As it can be seen, histogram bins regarding no motion have more counts than the rest. Similarly, in other corresponding CAMHIDs, only camera motion information is incorporated. Although CAMHID is capable of describing the camera motion efficiently, it has a limitation to represent the camera distance of the cinematographic directing semantic classes. As mentioned, camera distance is another important characteristics to be considered for classifying cinematographic shots. In this section, we extend the features representing the depth to overcome that limitation. The additional set of features is extracted from the readily available MVI. As the corresponding MBs of MVIs roughly represent the regions which preserve



(a) static shot



(b) contextual tracking shot

(c) focus tracking shot

(d) focus in shot

(e) focus out shot

(f) establishment shot

(g) chaotic shot

Figure 6.2.1: Cinematographic directing semantic shot classes and the corresponding CAMHID are shown in (a) - (g). The left column represents the first frames of the shots. Second column shows an intermediate frame of each shot. The third column represents the last frame of each shot. The right column shows the camera motion histogram descriptor by combining the local camera motion features on the four corners. The four corner local regions are identified on the last frame in (a) and the corresponding local histograms are identified in the corresponding CAMHID.

the camera movement, computing the ratio of the local MVI regions estimates the rough local depth. In order to do that, for each $MVF^{i}_{(p,q)}$, $\forall p \in \{1, 2, 3\}$, $\forall q \in \{1, 2, 3\}$ and $\forall i \in \{1, 2, \ldots, m\}$, we count the number of MVIs belonging to each local region ($mvi^{i}_{(x,y)} \in MVF^{i}_{(p,q)}$). Formally, we write:

$$C_{(p,q)} = \sum mvi^{i}_{(x,y)} \in MVF^{i}_{(p,q)} \tag{6.1}$$

where, $C_{(p,q)}$ is the count of the number of MVI present in the local region (p, q) for the entire shot. Then, local counts are normalised. Formally, we write as follows.

$$\hat{C}_{(p,q)} = C_{(p,q)}/(v * t) \tag{6.2}$$

where, v is the number of possible motion vectors in a video frame and t is number of frames in an input video. The normalised features along with CAMHID is the feature vector used for classifying directing semantic classes of cinematographic shots. In the next section, we show the effectiveness of our proposed features.

6.3 Experimental Results

To show the performance of the proposed camera motion histogram descriptor and its extension, the features are experimented and evaluated in this section. To do that, we evaluate the classification performance on directing semantic classes of cinematographic shots using a dataset. Then, we compare the results with the state-of-the-art methods available in the literature. The dataset consists of training and testing sets and the performance is evaluated based on the precision, recall and

f_1-scores on the testing datasets. The following subsections describe the detail of evaluation procedure.

6.3.1 Dataset Preparation and Feature Extraction

To evaluate the classification performance of the proposed CAMHID and its extension, we conduct experiments on our own created dataset.

The dataset (Dataset 3) is created based on the directing semantic classes of cinematographic shot. Similar to the Datasets created in Chapter 5, we label a training set and a testing set manually. For training the SVM classifiers, we create training data from five Hollywood films (Mission Impossible II, The proposal, The Mummy Returns, Hotel Rwanda and The Terminal). Then, for testing the classification performance, a testing dataset is created. The shots are taken from five Hollywood films. The numbers of shots taken from different movies are given in Table 6.3.1 and detailed breakdown of each shot type in the testing data is given in Table 6.3.2.

Table 6.3.1: Detail of testing data in Dataset 3

Film Title	**Number of Shots Taken**	**Total Duration**
The King's Speech	1853	118 min
Lord of The Rings I	1538	99 min
Kids are Alright	717	48 min
Mission Impossible II	564	33 min
A Beautiful Mind	886	58 min

Table 6.3.2: Detailed breakdown of the testing data in Dataset 3.

	S	**CT**	**FT**	**FI**	**FO**	**E**	**C**
no. of shots	2110	452	912	190	39	180	1675
'%'	37.96	8.13	16.41	3.42	0.70	3.24	30.14

6.3.2 SVM Classification

As in Chapter 5, In this chapter, we use One-Against-One SVM technique for cinematographic shot classification. To summarise the One-Against-One multi-class SVM, let us assume that we have n training data in d dimensional space belonging to c classes $\{\mathbf{x}^i, y^i\}, \quad \mathbf{x}^i \in \mathbf{R}^d, \quad i = 1, \ldots, n, \quad y^i \in \{1, \ldots, c\}$. This approach constructs $c(c-1)/2$ classifiers using the training data. Each of the classifier is obtained by using the training data of the corresponding two classes. For detail explanation of SVM classification used in this chapter, please refer to SVM classification in Chapter 5. In this chapter, kernel selection is made experimentally. In the experiments, we consider three kernels, namely polynomial, sigmoid and RBF kernels. For each of the kernels, precision and recall rates are measured in 3-fold and 5-fold cross validation settings. Table 6.3.3 shows the experimental results. As it can be seen, RBF kernel turns out to be the best performer in this experiment. Therefore, we select RBF kernel for conducting the rest of experiments. The accuracy of SVM classification depends on the values of two parameters C and γ. Careful selection of these two parameters is important. Otherwise, the classifier may perform poorly in the testing phase. A cross-validation approach is commonly used to determine the best parameters. We find the best penalty pa-

rameter C from the range $\{2^{-5}, 2^{-4}, \ldots, 2^{10}\}$ and width control parameter γ from the range $\{2^{-10}, 2^{-1}, \ldots, 2^{5}\}$.

Table 6.3.3: Classification accuracy measurements on Dataset 1 of Chapter 5 using different kernels for SVM classification. For this experiment we set $\gamma = 2^{-8}$ and $C = 2^{9}$.

Kernels	**3-fold cross validation**		**5-fold cross validation**	
	precision (%)	recall (%)	precision (%)	recall(%)
Polynomial kernel (2nd order), $(\gamma x_i^T x_j + r)^d, r = 1$	91.33 ± 2.39	90.14 ± 1.82	92.83 ± 1.66	88.11 ± 2.95
Polynomial kernel (4th order)	92.51 ± 1.77	90.33 ± 1.82	91.21 ± 2.01	91.35 ± 2.10
Sigmoid kernel, $\tanh(\gamma x_i^T xj + r), r = 1$	88.87 ± 1.62	82.48 ± 2.72	89.05 ± 1.21	86.72 ± 2.99
RBF kernel, $e^{-\gamma \|x^i, x^j\|^2}$	92.99 ± 1.26	94.07 ± 1.09	93.02 ± 1.39	94.34 ± 0.97

6.3.3 Evaluation

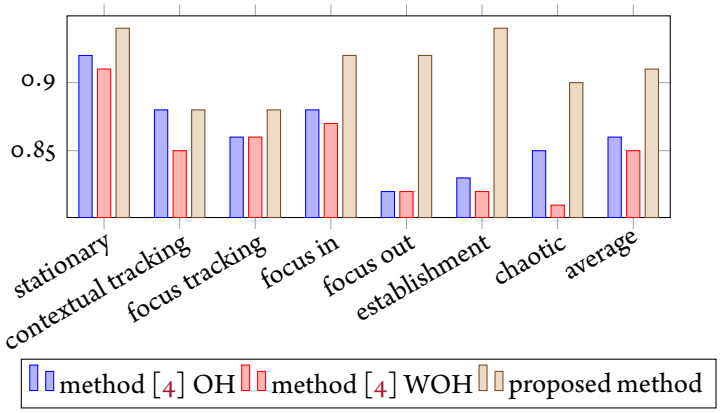
The performance of CAMHID descriptor and cinematographic features are evaluated in this subsection. The effectiveness is shown by using confusion matrix and by comparing recall rates, precision rates and f_1 scores.

The performance of shot classification is evaluated using Dataset 3 to show the ability of CAMHID and extended features in classifying cinematographic shots into the directing semantic classes. After training the SVM classifier using the training data of Dataset 3, the performance is evaluated using the testing data of Dataset 3. The confusion matrix is reported in Table 6.3.4. It is found that stationary shots are mostly confused with chaotic shots. This kind of misclassification mainly happen due to the threshold applied. This happens due to a small magnitude differences which fall near the borderline of motion magnitude. However, the amount of wrong classifications is at a minimum level. Focus tracking shots and contextual tracking shots introduce another level of confusion. Since the establishment shots also have similar motion patterns, this shot type also introduces additional confusion in classification. Although there is a level of confusion, the classification results using the proposed method are still promising. Table 6.3.5 shows the detailed classification performance using recall rates, precision rates and f_1-scores. To evaluate the performance of our proposed method, we compare the results with state-of-the-art methods described in [4]. Figures 6.3.1 shows the performance comparison of our method with the methods described in [4]. In [4], the authors proposed two methods to classify cinematographic shots into directing semantic classes. The first method classifies the shots with occlusion handling

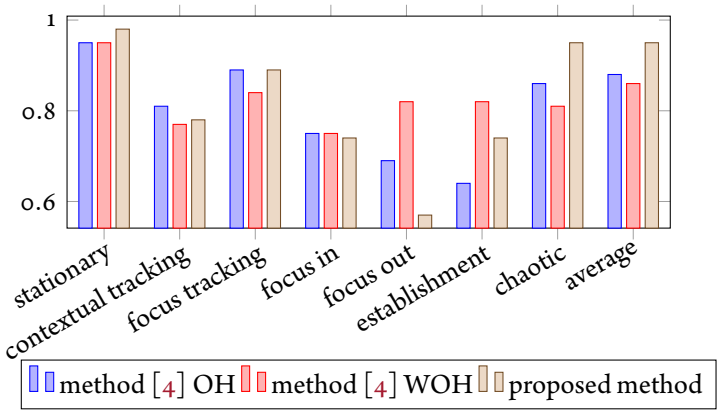
(OH) mechanism and the second method does so without occlusion handling (WOH) mechanism. As it can be seen in Figure 6.3.1 (a), the recall rate of our method for all the classes is higher than the results using the state-of-the-art approach. Although for most of the classes, precision rates using our method are higher than those using the method shown in [4], for other classes our precision rates show much lower rates than the state-of-the-art. To show a fairer comparison, Figure 6.3.1(c) demonstrates the comparison results of f_1-scores. It is shown that, the proposed method has higher f_1-scores for the classes except contextual tracking and focus-out classes. The average f_1-scores of [4] with occlusion handling and without occlusion handling are 83.03% and 81.55% respectively. However, it turns out that the average f_1-score of the proposed method is 85.02% which is higher than the other two methods.

Table 6.3.4: Confusion matrix of shot classification using Dataset 3.

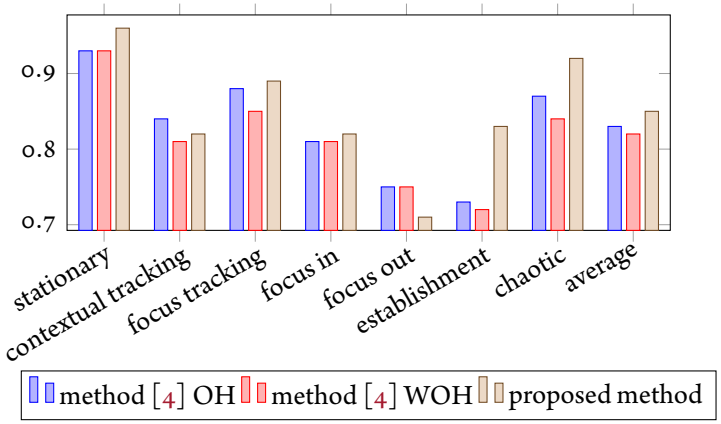
	S	**CT**	**FT**	**FI**	**FO**	**E**	**C**
S	94.22	0.57	0.71	0.57	0.28	0.57	3.08
CT	0.44	87.61	5.53	1.55	0.88	3.10	0.88
FT	0.88	8.00	88.16	0.66	0.55	0.66	1.10
FI	0.53	2.11	1.05	91.58	0.53	3.16	1.05
FO	0.00	2.56	2.56	0.00	92.31	0.00	2.56
E	0.00	2.22	0.56	1.67	0.00	94.44	1.11
C	1.55	1.07	3.04	2.03	0.66	1.31	90.33



(a) recall



(b) precision



(c) f_1-score

Figure 6.3.1: Cinematographic shot classification results comparison with [4] (a) Recall rate comparison, (b) Precision rate comparison, (c) f_1-score comparison.

Table 6.3.5: Recall (R), precision (P) and f_1 score (f_1) measures of shot classification performance using Dataset 3.

	S	**CT**	**FT**	**FI**	**FO**	**E**	**C**
R	94.22	87.17	88.16	88.95	89.74	94.44	90.33
P	98.08	77.87	89.23	72.53	55.56	73.91	94.68
f_1	96.11	82.26	88.69	79.91	68.63	82.92	92.45

6.4 Summarisation

In this chapter, the application of CAMHID features is experimented using the features set. The camera motion has been characterised by analysing the extracted raw motion vectors in the temporal domain. The temporal characterisation of the camera motions is then described by using a histogram, which combines local camera motion characterisation features. We have applied the proposed technique to classify cinematographic shots by extending the feature space. In the feature extension part, we consider the depth of the scene which is considered as one of the most important characteristics in cinematographic shot directing semantic classes. We have applied the motion descriptor with the extended feature on a separate dataset where shots are to be classified into directing semantic classes. We have evaluated and compared the performance of the proposed descriptor with state-of-the-art approaches. It has been demonstrated that the proposed descriptor has a strong capability to effectively discriminate different types of camera movements and shot types.

The pessimist sees difficulty in every opportunity. The optimist sees the opportunity in every difficulty.

Winston Churchill

7 Video Copy Detection Using CAMHID

7.1 Introduction

Recently, video copyright infringement is an important issue and identifying video copy has become an important research topic. Copyright video materials are often duplicated without permission. Duplicate videos are created using different transformations and/or pixel modification methods. Generally, duplicate videos are recognisable. In order to detect a copied video in a video database, features are extracted from a query video and passed to a copy detection algorithm to identify the copied video. In this chapter, we demonstrate the capability of CAMHID for video copy detection. The overall procedure is taken place in two steps: 1) identifying videos having similar motion contents using the CAMHID

using Earth Mover's Distance (EMD); 2) detecting video copies from the videos identified in step 1 through motion characteristics matching using Hamming distance.

The proposed technique is a robust method to detect video copies. It also works efficiently by narrowing down the search space significantly. First of all, the searching space is narrowed down by classifying the query videos into one of the basic camera motion classes. Then, the searching space is further narrowed down by measuring motion pattern similarity using EMD. EMD is well recognised for its robustness to measure the similarity between two distributions. Finally, the temporal similarities between the query video and the candidates are measured using Hamming distance.

The contributions of this chapter are as follows.

- We propose a novel motion content based similarity measuring technique to retrieve most similar videos in video databases.
- An efficient CAMHID based video copy detection technique is proposed to detect video copies in video databases.

The state-of-the-art of video copy detection is discussed in section 7.2. After that, the detail of the proposed techniques and experimental results are described in the following sections.

7.2 Background

Content based video copy detection works can be summarised into two broad categories: fingerprint based and sequence matching based methods. The fingerprint based techniques mainly extract a fingerprint to represent a video sequence. Fingerprints or video sequence descriptors are used to identify copied video sequences by measuring the similarity between video fingerprints. Sequence matching techniques match temporal signatures to identify copied video sequences.

Fingerprints are widely used for content based image retrieval which is also being applied for video copy detection. Such fingerprints are extracted from the video frames or from a keyframe of a video [155–158]. Fingerprint extraction for video copy detection can be classified into four categories, namely colour space based, temporal, spatial, and spatio-temporal feature based fingerprints [159]. Colour space based fingerprints are based on the statistics of the different colour spaces. Mostly, they are represented using histograms of colours of a region within a video frame. Since the colour of a video can be easily modified without being detected easily, this approach is not popular in practice. Moreover, this approach cannot be applied for the copy detection of gray-scale videos. In many applications, instead of colour, the luminance of a video frame is used for feature extraction. Spatial fingerprints are subdivided into two types, namely global and local fingerprints. Global fingerprints contain global spatial properties of frames while local spatial fingerprints represent local information using different segmentation methods. In global fingerprint methods, it is poplar to apply ordinal measure based techniques

[160–162]. In [161], video frames were divided into four ($= 2 \times 2$) regions and are ranked based on their intensities. Then, the rank matrix was considered to be the fingerprint of a frame. A similar technique was proposed in [160]. Fingerprint based techniques are fast in computation but suffer from lack of robustness.

Temporal fingerprints represent the temporal characteristics of a video in a compact manner [163]. In [164], a statistics based dissimilarity measure was used between two video sequences. In [165], the longest common sub-sequence was adopted for measuring temporal matching between two video sequences. In [166], different distance measures were used for the descriptor. Sequence matching based temporal fingerprints are often suffer from high computational load and lacks applicability for online video copy detection applications.

Similarity measure varies depending to query types and/or applications of video retrieval. According to [13], similarity measures can be classified into feature matching, text matching and ontology based matching. Feature matching techniques measure the average distance between the low level features of two videos. Text matching is another way for measuring similarity. In this case, the concepts of a video are learned off-line. When a user presents a text query, the similarity of the concept and the query is measured for retrieval. A higher value of the similarity metric implies a closer similarity. The main disadvantage of this approach is that learning a huge number of concepts from video is a very difficult problem. Ontology based matching measures similarity using the ontology between semantic concepts or the semantic relationship between keywords.

7.3 Proposed Video Copy Detection Method

Video copy detection is a daunting task. In a real life video database, the total number of videos may exceed tens of millions and the amount of contents is ever increasing. In this section, the proposed video copy detection method is described. The proposed technique uses the CAMHID features described in Chapter 5. First of all, the CAMHID features are extracted from a query video in order to determine if it is a copied/pirated version of one of the videos in a video database. For any video in a video database, the CAMHID features and temporal motion characteristics are extracted offline and stored in a feature database. Efficient copy detection mechanisms should consider finding a video copy by reducing the searching space in a video database. In order to do that, firstly we classify a query video into one of the shot classes proposed in Chapter 5. By doing this, the initial searching space greatly reduced. Upon query, the CAMHID features are extracted from the query video to measure the similarity with the videos in the corresponding class. In order to measure similarity, we use EMD. If the EMD between the query video and a video in the corresponding class is less than a predefined threshold, then the matched video will be used for the next step of similarity matching. In the next step, the Hamming distance between the motion characteristics of the query video and the motion characteristics of the selected video is computed. If the Hamming distance is less than a threshold, then the query video is declared as a copy of the matched one. Figure 7.3.1 shows the flow diagram of the proposed method. The proposed method is described in the following. Firstly, we formulate the problem

formally. Then, the detail of the similarity measure is described in the subsequent subsections.

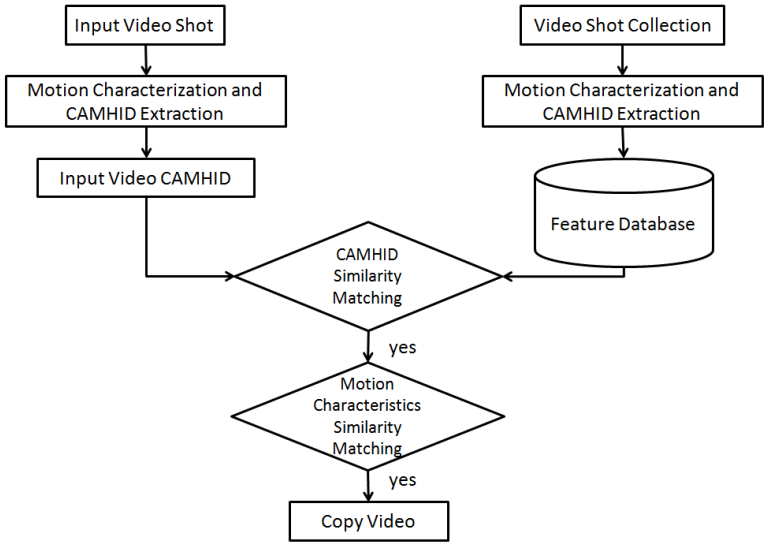


Figure 7.3.1: Framework for motion content similarity based video copy detection.

7.3.1 Problem Formulation

Let us assume that, we have a video sequence of m frames represented by $V = \{v_0, v_1, \ldots v_{m-1}\}$, where v_i represents i-th frame. A section of p frames of this video sequence is denoted by $V_{[j:j+p-1]}$, where j is the starting frame and $j+p-1$ indicates the ending frame satisfying $(j < m)$, $(j + p - 1 \leq m)$ and $V_{[j:j+p-1]} \subseteq V$.

Let us again assume that, we have a query video of n frames denoted by $Q = \{q_0, q_1, \ldots\ q_{n-1}\}$. We claim that Q is a copy of video V if the difference between Q and $V_{[j:j+p-1]}$, denoted by $\Delta(Q, V_{[j:j+p-1]})$, is at an acceptable level of dissimilarity, where $\Delta()$ is a viable distance measuring function.

7.3.2 Dissimilarity Measurements

As mentioned, when a query video is received by the copy detection system, firstly, the query video is classified based on the classification technique described in Chapter 5. Then, each video in the corresponding class has to go through two steps to measure the dissimilarity. Firstly, the searching space is further reduced by measuring the similarity between the query video's and target videos' CAMHID features. The similarity is measured using EMD. We use this technique because of its logical capability to measure the similarity between two distributions. Videos with acceptable levels of similarity are taken for further consideration in finding a video copy. Then, local motion characterisations are used to measure the temporal similarity between the query and each survived video. In this case, we measure the Hamming distance for finding the dissimilarity to declare the query video is a copy of the found video. In the following, we discuss about the earth mover distance and the Hamming distance measurements.

Earth Mover's Distance Based Similarity Matching

The earth mover's distance (EMD) [167] is used to measure the cross-bin distance between two distributions. It is a measure that computes the minimal cost to transform one distribution into another. In our proposed method, the similarity between the CAMHID of the query video and that of the target video is measured to assess the rough motion content similarity. We assume that the CAMHID feature of the query video is denoted by H_Q, and the CAMHID feature of the target video is denoted by H_V. We also assume that the length of the CAMHID feature is l. A

histogram H of length l is formally defined as follows.

$$H = \{(i, w_i) : 1 \leq i \leq l\}, \tag{7.1}$$

where, (i, w_i) indicates the count of the i-th bin is w_i. For measuring the transformation of a histogram to another histogram, the minimum flow is represented as follows.

$$F = \{(i, w_i; j, w_j) : (i, w_i) \in H, (j, w_j) \in H\}, \tag{7.2}$$

where, $(i, w_i; j, w_j)$ represents the bin weight flow from bin i to bin j.

Under the above representation, we want to convert H_Q into H_V. Therefore, the formal representations of the CAMHID features are as follows.

$$H_Q = \{i^q, w_i^q : (i^q, w_i^q) \in H_Q\}, \tag{7.3}$$

and

$$H_V = \{i^v, w_i^v : (i^v, w_i^v) \in H_V\}, \tag{7.4}$$

where, (i^q, w_i^q) and (i^v, w_i^v) indicate the count of the i^q-th and i^v bin are w_i^q and w_i^v respectively. We need to compute the minimum cost for transformation of histogram H_Q into histogram H_V to measure the motion content similarity. In order to do that, we normalise both of the histograms such that $\sum_i w_i^q = 1$ and

$\sum_i w_i^v = 1$. The EMD to transform H_Q into H_V is as follows.

$$EMD(H_Q, H_V) = \min_{T=\{t_{i^q,w_i^q;j^q,w_j^q}:(i^q,w_i^q;j^q,w_j^q)\in F\}} \sum t_{i^q,w_i^q;j^q,w_j^q}.d_{i^q,j^q} \tag{7.5}$$

$$s.t. \begin{cases} \sum_{(j^q,w_j^q)\in H_Q} t_{i^q,w_i^q;j^q,w_j^q} = H_{V_{i,w_i}} & \forall(i^q, w_i^q) \in H_Q \\ \sum_{(i^v,w_i^v)\in H_V} t_{i^v,w_i^v;j^v,w_j^v} = H_{Q_{j,w_i}} & \forall(j^v, w_j^v) \in H_V \\ t_{i,w_i;j,w_j} \geq 0 & \forall(i, w_i; j, w_j) \in F \end{cases} \tag{7.6}$$

where T is a set of t_{i,w_i,j,w_j} representing total flow from bin i to bin j and $d_{i,j}$ indicates the ground distance between bin i to bin j. The ground distance is defined as follows.

$$d_{i,j} = ||(i, w_i)^T - (j, w_j)^T||. \tag{7.7}$$

Based on the above definition of EMD, we use the following equation for taking a decision for further processing.

$$similar = \begin{cases} yes & EMD(H_Q, H_V) < \tau_1 \\ no & otherwise \end{cases} \tag{7.8}$$

If the above equation is satisfied, then the target video V is passed to the next level for a detailed similarity measure.

Hamming Distance Based Dissimilarity Measurements

After finding the similar video, the video copy is detected by computing the Hamming distance. The Hamming distance is a measure to estimate the similarity in the temporal motion characteristics. In order to do so, motion characteristics described in Chapter 5 is used. Let us assume that the oriented angle of the compact representation of the motion of a local region in a query video Q is denoted by $\{\theta^{Q}_{(x,y)}\}$, where $x \in 1, 2, 3$ and $y \in \{1, 2, 3\}$ represent the coordinates of the local region. Similarly, the oriented angle of the compact representation of the motion of a local region in a target video V is denoted by $\{\theta^{V}_{(x,y)}\}$, where $x \in \{1, 2, 3\}$ and $y \in \{1, 2, 3\}$ are the coordinates of the local region. The Hamming distance between query video Q and the target video $V_{[j:j+p-1]}$ is measured as follows.

$$sim(Q_{x,y}, V_{[j:j+p-1]_{x,y}}) = \begin{cases} 1 & \forall |\theta^{V_{[j:j+p-1]}}_{(x,y)} - \theta^{Q}_{(x,y)}| < \tau_2 \\ 0 & otherwise \end{cases} \tag{7.9}$$

If Eq. (7.9) is satisfied in a single local region, then it means that both video's local motion characteristics are the same. As each video is segmented into 9 ($=3 \times 3$) local regions, the similarities for all local regions are measured. Finally, we consider the query video is a copied one if the video contains similar motion contents in at least 7 out of 9 local regions.

7.4 Experimental Results

In order to measure the performance of the proposed video copy detection, in the following, a detailed experimental procedure is described.

7.4.1 Testing Dataset

We use TRECVID 2008 dataset for evaluating the performance of video copy detection. This dataset contains 200 hours of videos from Dutch television programmes. This dataset has been used for the TRECVID 2008 video a copy detection evaluation campaign. The dataset contains 134 positive query video clips and 67 negative query video clips. The query videos were transformed using ten different transformation methods, namely camcording (CAM), picture in picture (PIP), insertion of patterns (IoP), strong re-encoding (SRE), change of gamma (CG), photometric attacks (PA), geometric attacks (GA), 3 random transformations from 6/7 (3RT-6/7), 5 random transformations from 6/7 (5RT-6/7), and 5 random transformations (5RT). As a result, a good number of query videos were prepared to test the performance of the proposed video copy detection method. In our method, we assume that the videos are segmented into frames. In the feature database, the classified (i.e., static, pan, tilt and zoom) CAMHID features are stored for a fast similarity measurements. Upon query, the query video shot is classified and the similarity measurement is performed within the corresponding class.

7.4.2 Evaluation

In order to evaluate the quality of performance of the proposed method, the Normal Detection Cost Rates (NDCR) is used. NDCR integrates the cost of failing to detect true positive and cost of including false positive. The lower the value of NDCR is, the better the result is. A formal definition of NDCR can be found at [168] which is summarized as follows.

$$NDCR = \frac{FN}{TP} + \eta \times \frac{FP}{LH} \tag{7.10}$$

where FN, TP, FP and LH are false-negative, true-positive, false-positive and length in hour respectively. η is a constant. In our implementation, *eta* is set to 0.5. Figure 7.4.1 shows the performance of our proposed method in comparison to the method proposed in [169] and with the best results achieved in the TRECVID 2008 competition. We have not compared the results for picture in picture modificationas we have not addressed this problem in our proposed method. As it can be seen in Figure 7.4.1, for photometric attack, our proposed method outperforms the compared methods. For the change of gamma modification, our proposed method is performing as good as the best result achieved in TRECVID 2008. For other modifications, although our method cannot beat the best result in TRECVID 2008, our proposed method is performing a lot better than the method proposed in [169].

Although the proposed method could not beat all of the state-of-the-art methods, our proposed method has added advantages. The proposed method is con-

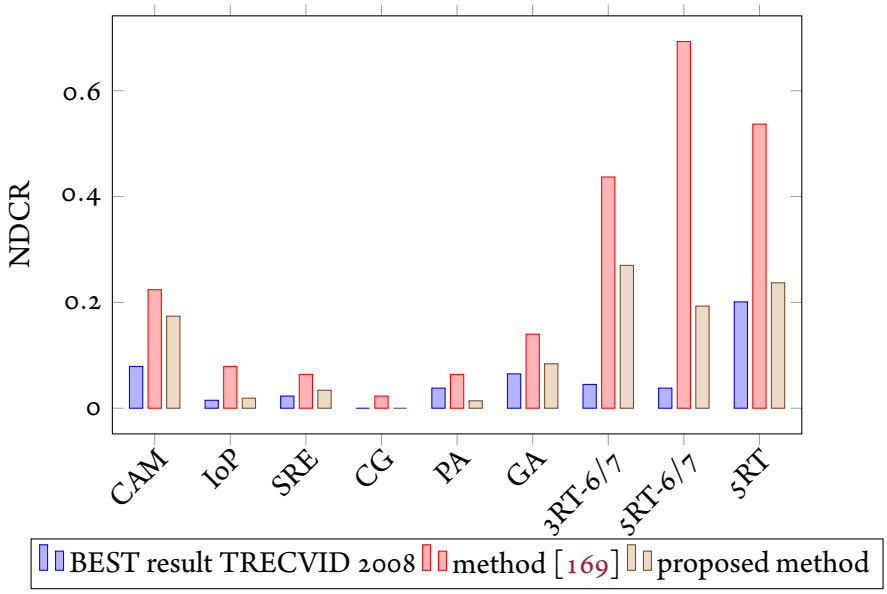


Figure 7.4.1: Video copy detection result comparison with the state of the art methods.

sidered to be a fast solution for video copy detection task. As the searching space is greatly reduced by using the camera motion classification and the CAMHID similarity measurement, the proposed method can be used in a very large scale database.

7.5 Summarisation

In this chapter, we have proposed a method to detect video copy from a video database. Firstly, searching space is reduced by classifying video into one of the basic camera motion calsses. Then, searching space is further narrowed by measuring the similarity between respective video CAMHID features. In order to measure the CAMHID similarity, we have used the earth mover's distance. In the final step,

we go through a detailed similarity measurement using temporal motion characteristics. At this stage, we have measured the Hamming distance to identify a video copy. According to our experiments, we have found that our proposed method is a competitive one. Although our proposed method has not outperformed the compared method, our method is considered as a fast video copy detection method for a large video database.

Do the difficult things while they are easy and do the great things while they are small. A journey of a thousand miles must begin with a single step.

Lao Tzu

8 Conclusion

8.1 Conclusion

Cinematographic shot classification is a vital and a challenging task due to various movie genres, different shooting techniques and many more shot types than other video domain. Identifying motion content is useful for shot level semantic analysis. Moreover, indexing such video shots in video databases may provide efficient solutions to many applications, such as semantic understanding of movies, automatic movie editing for theme representation, constructing movie structure for browsing movies, video summarisation, and movie genre classification. Although movie shot classification has many potential applications, it has not been addressed adequately. We have proposed three cinematographic shot

indexing methods in this thesis. Aside from that, we have proposed a video copy retrieval technique based on the motion descriptor technique proposed in this thesis. The summary of the thesis is presented in the following section.

8.2 Summary

In Chapter 2, a detailed review of the related work is presented. At the beginning, the detailed of the background and the generic state-of-the-art shot classification were discussed. Then, the detail of the cinematographic shots classification task was discussed.

In Chapter 3, we introduce the context saliency to measure the visual attention distributed in keyframes for movie shot classification. Different from traditional saliency maps, the context saliency map is generated by removing redundancy from contrast saliency and incorporating geometry constrains. The context saliency is later combined with colour and texture features to generate feature vectors. Support Vector Machine (SVM) is used to classify keyframes into predefined shot classes. Different from the existing work of either performing in a certain movie genre or classifying movie shot into limited directing semantic classes, the proposed method has three unique features: 1) the context saliency significantly improves movie shot classification; 2) our method works for all movie genres; 3) our method deals with the most common types of video shots in movies. The experimental results indicate that the proposed method is effective and efficient for movie shot classification.

In Chapter 4, Among many video types, movie content indexing and retrieval

are significantly challenging tasks because of the wide variety of shooting techniques and the broad range of genres. A movie consists of a series of video shots. Managing a movie at shot level provides a feasible way for movie understanding and summarisation. Consequently, an effective shot classification is greatly desired for advanced movie management. We explore novel domain specific features for effective shot classification. Experimental results show that the proposed method classifies movie shots from wide range of movie genres with improved accuracies compared to the existing work.

In Chapter 5, we propose a non-parametric camera motion descriptor called CAMHID for video shot classification. In the proposed method, a motion vector field (MVF) is constructed for each consecutive video frames by computing the motion vector of each macroblock (MB). Then, the MVFs are divided into a number of local region of equal size. Next, the inconsistent/noisy motion vectors of each local region are eliminated by a motion consistency analysis. The remaining motion vectors of each local region from n consecutive frames are further collected for a compact representation. Initially, a matrix is formed using the motion vectors. Then, the matrix is decomposed using the singular value decomposition (SVD) technique to represent the dominant motion. Finally, the angle of the most dominant principal component is computed and quantised to represent the motion of a local region by using a histogram. In order to represent the global camera motion, the local histograms are combined. The effectiveness of the proposed motion descriptor for video shot classification is tested by using support vector machine (SVM). Firstly, the proposed camera motion descriptors for video shots

classification are computed on a video dataset consisting of regular camera motion patterns (e.g., pan, zoom, tilt, static). We also show that our approach outperforms a state-of-the-art video shot classification method.

In Chapter 6, we apply the camera motion descriptors CAMHID along with an extended set of features to the classification of cinematographic shots. The experimental results show that the proposed shot level camera motion descriptor has a strong discriminative capability to classify different camera motion patterns of different videos effectively. We also show that our approach outperforms state-of-the-art methods.

In Chapter 7, the performance of CAMHID is evaluated in video copy detection and retrieval tasks. In order to to that, video shots are firstly classified into one of the camera motion classes. Then, a two step similarity measurement is conducted to identify and retrieve the video copy. The proposed method is robust in identifying video copy. Moreover, the speed of the proposed method makes it viable to be applied in large scale video databases for identifying video copies.

8.3 Future Work

In this thesis, we only consider the quality of motion and quantity of motion is not used at all. Quantity has also a good role to play in video indexing tasks. In future, the quantitative analysis of camera motions is to be done which will also be useful for a better camera motion descriptor. The proposed motion descriptor is useful for professionally captured videos. In future, we will extend the motion descriptor to form a generalised camera motion descriptor. Using such a descriptor, home

videos can also be indexed. It will be interesting to see the performance of the proposed descriptor in personalised video content analysis, video recommendation, and video summarisation. These are left for our future work. For video copy detection, our proposed method cannot handle the problem if a copied video is embedded in another video. In future, we will further investigate this problem.

References

[1] Min Xu, Jinqiao Wang, M.A. Hasan, Xiangjian He, Changsheng Xu, Hanqing Lu, and J.S. Jin. Using context saliency for movie shot classification. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 3653 –3656, sept. 2011.

[2] Chong-Wah Ngo, Ting-Chuen Pong, and Hong-Jiang Zhang. Motion analysis and segmentation through spatio-temporal slices processing. *Image Processing, IEEE Transactions on*, 12(3):341–355, 2003.

[3] Ralph Ewerth, Martin Schwalb, Paul Tessmann, and Bernd Freisleben. Estimation of arbitrary camera motion in mpeg videos. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 512–515. IEEE, 2004.

[4] Hee Lin Wang and Loong Fah Cheong. Taxonomy of directing semantics for film shot classification. *IEEE Trans. Circuits Syst. Video Techn.*, pages 1529–1542, 2009.

[5] Ramesh Jain. Teleexperience: communicating compelling experience. In *Proceedings of the ninth ACM international conference on Multimedia*, Multimedia '01, pages 1–1, New York, NY, USA, 2001. ACM.

[6] J. Steiff. *The Complete Idiot's Guide to Independent Filmmaking*. Complete Idiot's Guide to. Alpha, 2005.

[7] Steven Douglas Katz. *Film directing shot by shot: visualizing from concept to screen*. Gulf Professional Publishing, 1991.

[8] J. Cantine, S. Howard, and B. Lewis. *Shot by shot: a practical guide to filmmaking*. Pittsburgh Filmmakers, 2000.

[9] Zheng-Jun Zha, Meng Wang, Yan-Tao Zheng, Yi Yang, Richang Hong, and Tat-Seng Chua. Interactive video indexing with statistical active learning. *Multimedia, IEEE Transactions on*, 14(1):17–27, 2012.

[10] Jun Wu and Marcel Worring. Efficient genre-specific semantic video indexing. *Multimedia, IEEE Transactions on*, 14(2):291–302, 2012.

[11] Bo Geng, Yangxi Li, Dacheng Tao, Meng Wang, Zheng-Jun Zha, and Chao Xu. Parallel lasso for large-scale video concept detection. *Multimedia, IEEE Transactions on*, 14(1):55–65, 2012.

[12] Mats Sjoberg, Markus Koskela, Satoru Ishikawa, and Jorma Laaksonen. Real-time large-scale visual concept detection with linear classifiers. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 421–424. IEEE, 2012.

[13] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. A survey on visual content-based video indexing and retrieval. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 41(6):797–819, 2011.

[14] S. Dagtas, W. Al-Khatib, A. Ghafoor, and R.L. Kashyap. Models for motion-based video indexing and retrieval. *Image Processing, IEEE Transactions on*, 9(1):88 –101, jan 2000.

[15] Ba Tu Truong, S. Venkatesh, and C. Dorai. Discovering semantics from visualizations of film takes. In *Multimedia Modelling Conference, 2004. Proceedings. 10th International*, pages 109 – 116, jan. 2004.

[16] Frank Nack and Alan Parkes. The application of video semantics and theme representation in automated video editing. *Multimedia Tools Appl.*, 4(1):57–83, January 1997.

[17] Riad Hammoud and Roger Mohr. Interactive tools for constructing and browsing structures for movie films. In *IN ACM Multimedia*, pages 497–498, 2000.

[18] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. A generic framework of user attention model and its application in video summarization. *Multimedia, IEEE Transactions on*, 7(5):907–919, 2005.

[19] Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. Movie genre classification via scene categorization. In *Proceedings of the international conference on Multimedia*, MM '10, pages 747–750, New York, NY, USA, 2010. ACM.

[20] Shih-Fu Chang, William Chen, Horace J Meng, Hari Sundaram, and Di Zhong. A fully automated content-based video search engine supporting spatiotemporal queries. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):602–615, 1998.

[21] C. Dorai and S. Venkatesh. *Media Computing: Computational Media Aesthetics*. The Kluwer International Series in Video Computing / Editor Mubarak Shah. Kluwer, 2002.

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

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

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

[25] Keewon Seo, Jaeseung Ko, Ilkoo Ahn, and Changick Kim. An intelligent display scheme of soccer video on mobile devices. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(10):1395 –1401, oct. 2007.

[26] H.M. Zawbaa, N. El-Bendary, A.E. Hassanien, and A. Abraham. Svm-based soccer video summarization system. In *Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on*, pages 7 –11, oct. 2011.

[27] Lexing Xie, Peng Xu, Shih fu Chang A, Ajay Divakaran, and Huifang Sun B. Structure analysis of soccer video with hidden markov models. In *Pattern Recognition Letters*, pages 767–775, 2002.

[28] Ling yu Duan, Min Xu, Xiao dong Yu, and Qi Tian. A unified framework for semantic shot classification in sports video. *Transactions on Multimedia*, 7:2005, 2002.

[29] A. Ekin and A.M. Tekalp. Robust dominant color region detection and color-based applications for sports video. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I – 21–4 vol.1, sept. 2003.

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

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

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

[33] Eddie Cooke, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Gareth JF Jones, Hervé Le Borgne, Hyowon Lee, Seán Marlow, Kieran McDonald, Mike McHugh, et al. Trecvid 2004 experiments in dublin city university. 2004.

[34] Yihong Gong, Lim Teck Sin, Chua Hock Chuan, Hongjiang Zhang, and Masao Sakauchi. Automatic parsing of tv soccer programs. In *Multimedia Computing and Systems, 1995., Proceedings of the International Conference on*, pages 167–174. IEEE, 1995.

[35] Antonio Torralba and Aude Oliva. Depth estimation from image structure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1226–1238, September 2002.

[36] T. Nagai, T. Naruse, M. Ikehara, and A. Kurematsu. Hmm-based surface reconstruction from single images. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 2, pages II–561 – II–564 vol.2, 2002.

[37] Golnaz Abdollahian, Cüneyt M Taskiran, Zygmunt Pizlo, and Edward J Delp. Camera motion-based analysis of user generated video. *Multimedia, IEEE Transactions on*, 12(1):28–41, 2010.

[38] JungHwan Oh and P. Sankuratri. Automatic distinction of camera and object motions in video sequences. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 81–84 vol.1.

[39] Rong Jin, Yanjun Qi, and A. Hauptmann. A probabilistic model for camera zoom detection. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 859–862 vol.3.

[40] Sangkeun Lee and III Hayes, M.H. Real-time camera motion classification for content-based indexing and retrieval using templates. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–3664–IV–3667, May.

[41] Ling-Yu Duan, J.S. Jin, Qi Tian, and Chang-Sheng Xu. Nonparametric motion characterization for robust classification of camera motion patterns. *Multimedia, IEEE Transactions on*, 8(2):323–340, April.

[42] Dong-Jun Lan, Yu-Fei Ma, and Hong-Jiang Zhang. A systemic framework of camera motion analysis for home video. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I–289–92 vol.1, Sept.

[43] Jae-Gon Kim, Hyun Sung Chang, Jinwoong Kim, and Hyung-Myung Kim. Efficient camera motion characterization for mpeg video indexing. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, pages 1171–1174. IEEE, 2000.

[44] Anil K Jain, Aditya Vailaya, and Xiong Wei. Query by video clip. *Multimedia systems*, 7(5):369–384, 1999.

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

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

[47] Minh-Son Dao, FGB DeNatale, and Andrea Massa. Video retrieval using video object-trajectory and edge potential function. In *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, pages 454–457. IEEE, 2004.

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

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

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

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

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

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

[54] Sylvie Jeannin and Ajay Divakaran. Mpeg-7 visual motion descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):720–724, 2001.

[55] Patrick Bouthemy, Marc Gelgon, and Fabrice Ganansia. A unified approach to shot change detection and camera motion characterization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(7):1030–1044, 1999.

[56] Chong-Wah Ngo, Ting-Chuen Pong, and Hong-Jiang Zhang. Motion-based video representation for scene change detection. *International Journal of Computer Vision*, 50(2):127–142, 2002.

[57] Chong-Wah Ngo, Ting-Chuen Pong, Hong-Jiang Zhang, and R.T. Chin. Motion-based video representation for scene change detection. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 827–830 vol.1.

[58] Michal Irani and P Anandan. Video indexing based on mosaic representations. *Proceedings of the IEEE*, 86(5):905–921, 1998.

[59] Yap-Peng Tan, Drew D Saur, Sanjeev R Kulkami, and Peter J Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(1):133–146, 2000.

[60] Zoran Duric and Azriel Rosenfeld. Image sequence stabilization in real time. *Real-Time Imaging*, 2(5):271–284, 1996.

[61] Yi-Sheng Yao and Rama Chellappa. Electronic stabilization and feature tracking in long image sequences. Technical report, DTIC Document, 1995.

[62] Nilesh V Patel and Ishwar K Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30(4):583–592, 1997.

[63] Wei Xiong and John Chung-Mong Lee. Efficient scene change detection and camera motion annotation for video classification. *Computer Vision and Image Understanding*, 71(2):166–181, 1998.

[64] Yu-Fei Ma and Hong-Jiang Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381. ACM, 2003.

[65] T. Zhang and C.-C.Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *Speech and Audio Processing, IEEE Transactions on*, 9(4):441 –457, May 2001.

[66] Zhu Liu, Yao Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene segmentation and classification. *J. VLSI Signal Process. Syst.*, 20(1/2):61–79, October 1998.

[67] D. Brezeale and D.J. Cook. Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):416 –430, May 2008.

[68] Rainer Lienhart and Wolfgang Effelsberg. Automatic text segmentation and text recognition for video indexing. *Multimedia Systems.*, 8(1):69–81, January 2000.

[69] Radu S Jasinschi, Nevenka Dimitrova, Thomas McGee, Lalitha Agnihotri, John Zimmerman, and Dongge Li. Integrated multimedia processing for topic segmentation and classification. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 3, pages 366–369. IEEE, 2001.

[70] Nevenka Dimitrova, Lalitha Agnihotri, and Gang Wei. Video classification based on hmm using text and faces. In *European Conference on Signal Processing*. Citeseer, 2000.

[71] Polyxeni Katsiouli, Vassileios Tsetsos, and Stathes Hadjiefthymiades. Semantic video classification based on subtitles and domain terminologies. In *KAMC*, 2007.

[72] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. 2013.

[73] Bilge Gunsel, A Mufit Ferman, and A Murat Tekalp. Video indexing through integration of syntactic and semantic features. In *Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on*, pages 90–95. IEEE, 1996.

[74] Kim Shearer, Chitra Dorai, and Svetha Venkatesh. Incorporating domain knowledge with video and voice data analysis in news broadcasts. In *MDM/KDD*, pages 46–53, 2000.

[75] Marco Bertini, Alberto Del Bimbo, and Pietro Pala. Content-based indexing and retrieval of tv news. *Pattern Recognition Letters*, 22(5):503–516, 2001.

[76] Alan Hanjalic, Geerd Kakes, Reginald L Lagendijk, and Jan Biemond. Dancers: Delft advanced news retrieval system. In *Proceedings of SPIE*, volume 4315, page 301, 2001.

[77] Ichiro Ide, Koji Yamamoto, and Hidehiko Tanaka. Automatic video indexing based on shot classification. In *Advanced Multimedia Content Processing*, pages 87–102. Springer, 1999.

[78] Stefan Eickeler and Stefan Muller. Content-based video indexing of tv broadcast news using hidden markov models. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 2997–3000. IEEE, 1999.

[79] Peng Xu, Lexing Xie, Shih-Fu Chang, Ajay Divakaran, Anthony Vetro, and Huifang Sun. Algorithms and system for segmentation and structure analysis in soccer video. In *ICME*, volume 1, pages 928–931. Citeseer, 2001.

[80] Alexander G Hauptmann and Michael J Witbrock. Story segmentation and detection of commercials in broadcast news video. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on*, pages 168–179. IEEE, 1998.

[81] Rainer Lienhart, Christoph Kuhmunch, and Wolfgang Effelsberg. On the detection and recognition of television commercials. In *Multimedia Computing and Systems' 97. Proceedings., IEEE International Conference on*, pages 509–516. IEEE, 1997.

[82] Ba Tu Truong and Chitra Dorai. Automatic genre identification for content-based video categorization. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 230–233. IEEE, 2000.

[83] Vikrant Kobla, Daniel DeMenthon, and David S Doermann. Identifying sports videos using replay, text, and camera motion features. In *Electronic Imaging*, pages 332–343. International Society for Optics and Photonics, 1999.

[84] Jincheng Huang, Zhu Liu, Yao Wang, Yu Chen, and Edward K Wong. Integration of multimodal features for video scene classification based on hmm. In *Multimedia Signal Processing, 1999 IEEE 3rd Workshop on*, pages 53–58. IEEE, 1999.

[85] Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. Automatic recognition of film genres. In *ACM multimedia*, volume 95, pages 295–304, 1995.

[86] Emile Sahouria and Avideh Zakhor. Content analysis of video using principal components. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(8):1290–1298, 1999.

[87] Niels Haering, Richard J Qian, and M Ibrahim Sezan. A semantic event-detection approach and its application to detecting hunts in wildlife video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 10(6):857–868, 2000.

[88] Simon Moncrieff, Chitra Dorai, and Svetha Venkatesh. Detecting indexical signs in film audio for scene interpretation. In *ICME*, 2001.

[89] Hao Pan, P Van Beek, and MI Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 3, pages 1649–1652. IEEE, 2001.

[90] Ba Tu Truong and Svetha Venkatesh. Determining dramatic intensification via flashing lights in movies. In *ICME*, 2001.

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

[92] Hisashi Miyamori and S-I Iisaku. Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 320–325. IEEE, 2000.

[93] G Sudhir, John Chung-Mong Lee, and Anil K Jain. Automatic classification of tennis video for high-level content-based retrieval. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pages 81–90. IEEE, 1998.

[94] Di Zhong and Shih-Fu Chang. Structure analysis of sports video using domain models. In *IEEE ICME*. Citeseer, 2001.

[95] A Bonzanini, Riccardo Leonardi, and Pierangelo Migliorati. Event recognition in sport programs using low-level motion indices. In *ICME*, 2001.

[96] Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for tv baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 105–115. ACM, 2000.

[97] Drew D Saur, Yap-Peng Tan, Sanjeev R Kulkarni, and Peter J Ramadge. Automated analysis and annotation of basketball video. In *Electronic Imaging'97*, pages 176–187. International Society for Optics and Photonics, 1997.

[98] Wensheng Zhou, Asha Vellaikal, and CC Kuo. Rule-based video classification system for basketball video indexing. In *Proceedings of the 2000 ACM workshops on Multimedia*, pages 213–216. ACM, 2000.

[99] Ling-Yu Duan, Min Xu, Tat-Seng Chua, Qi Tian, and Chang-Sheng Xu. A mid-level representation framework for semantic sports video analysis. In *Proceedings of the eleventh ACM international conference on Multimedia*, MULTIMEDIA '03, pages 33–44, New York, NY, USA, 2003. ACM.

[100] Yap peng Tan, Drew D. Saur, Sanjeev R. Kulkarni, and Peter J. Ramadge. Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Trans. on Circuits and Systems for Video Technology*, 10:133–146, 1998.

[101] Lexing Xie, Peng Xu, Shih-Fu Chang, Ajay Divakaran, and Huifang Sun. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recogn. Lett.*, 25(7):767–775, May 2004.

[102] Chong-Wah Ngo, Ting-Chuen Pong, and Hong-Jiang Zhang. On clustering and retrieval of video shots. In *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA '01, pages 51–60, New York, NY, USA, 2001. ACM.

[103] F.M. Idris and S. Panchanathan. Spatio-temporal indexing of vector quantized video sequences. *Circuits and Systems for Video Technology, IEEE Transactions on*, 7(5):728 –740, oct 1997.

[104] JungHwan Oh, Maruthi Thenneru, and Ning Jiang. Hierarchical video indexing based on changes of camera and object motions. In *Proceedings of*

the 2003 ACM symposium on Applied computing, SAC '03, pages 917–921, New York, NY, USA, 2003. ACM.

[105] Paul Over, Tzveta Ianeva, Wessel Kraaij, and Alan F. Smeaton. Trecvid 2005 - an overview. In *In Proceedings of TRECVID 2005*, 2005.

[106] Yu-Hsuan Ho, C.-W. Lin, Jing-Fung Chen, and H.-Y.M. Liao. Fast coarse-to-fine video retrieval using shot-level spatio-temporal statistics. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(5):642 – 648, May 2006.

[107] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga. Sports video categorizing method using camera motion parameters. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 1*, ICME '03, pages 461–464, Washington, DC, USA, 2003. IEEE Computer Society.

[108] M. Lazarescu and S. Venkatesh. Using camera motion to identify types of american football plays. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 1*, ICME '03, pages 181–184, Washington, DC, USA, 2003. IEEE Computer Society.

[109] Mei Han, Wei Hua, Wei Xu, and Yihong Gong. An integrated baseball digest system using maximum entropy method. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, pages 347–350, New York, NY, USA, 2002. ACM.

[110] Jeroen Vendrig and Marcel Worring. Systematic evaluation of logical story unit segmentation. *Multimedia, IEEE Transactions on*, 4(4):492–499, 2002.

[111] A Aydin Alatan, Ali N Akansu, and Wayne Wolf. Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools and applications*, 14(2):137–151, 2001.

[112] Jeho Nam, Masoud Alghoniemy, and Ahmed H Tewfik. Audio-visual content-based violent scene characterization. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, volume 1, pages 353–357. IEEE, 1998.

[113] Caterina Saraceno and Riccardo Leonardi. Identification of story units in audio-visual sequences by joint audio and video processing. In *Image*

Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, volume 1, pages 363–367. IEEE, 1998.

[114] Minerva M Yeung and Boon-Lock Yeo. Video content characterization and compaction for digital library applications. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 45–58, 1997.

[115] Hee Lin Wang and Loong Fah Cheong. Film shot classification using directing semantics. In *ICPR'08*, pages 1–4, 2008.

[116] Shuhui Wang, Shuqiang Jiang, Qingming Huang, and Wen Gao. Shot classification for action movies based on motion characteristics. In *ICIP'08*, pages 2508–2511, 2008.

[117] Luca Canini, Sergio Benini, and Riccardo Leonardi. Classifying cinematographic shot types. *Multimedia Tools and Applications*, pages 1–23. 10.1007/s11042-011-0916-9.

[118] I. Cherif, V. Solachidis, and I. Pitas. Shot type identification of movie content. In *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, pages 1 –4, feb. 2007.

[119] R. Kindermann, J.L. Snell, and American Mathematical Society. *Markov random fields and their applications*. Contemporary mathematics. American Mathematical Society, 1980.

[120] S. Bhattacharya, R. Mehran, R. sukthankar, and M. Shah. Classification of cinematographic shots using lie algebra and its application to complex event recognition. *Multimedia, IEEE Transactions on*, PP(99):1–1, 2014.

[121] Linjun Yang, Bo Geng, Alan Hanjalic, and Xian-Sheng Hua. Contextual image retrieval model. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 406–413. ACM, 2010.

[122] Liu Huiying, Xu Min, Huang Qingming, Jin Jesse Sheng, Jiang Shuqiang, and Xu Changsheng. A close-up detection method for movies. 2010.

[123] Liang Shi, Jinqiao Wang, Lei Xu, Hanqing Lu, and Changsheng Xu. Context saliency based image summarization. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 270–273. IEEE, 2009.

[124] Momotaz Begum and Fakhri Karray. Visual attention for robotic cognition: a survey. *Autonomous Mental Development, IEEE Transactions on*, 3(1):92–105, 2011.

[125] Gustavo Deco and Tai Sing Lee. A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomputing*, 44:775–781, 2002.

[126] Gustavo Deco and Edmund T Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44(6):621–642, 2004.

[127] Linda J Lanyon and Susan L Denham. A model of active visual search with object-based attention guiding scan paths. *Neural Networks*, 17(5):873–897, 2004.

[128] Peter F Dominey and Michael A Arbib. A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, 2(2):153–175, 1992.

[129] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial intelligence*, 78(1):507–545, 1995.

[130] Yaoru Sun and Robert Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.

[131] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[132] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[133] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012.

[134] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[135] Feng Liu and Michael Gleicher. Region enhanced scale-invariant saliency detection. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1477–1480. IEEE, 2006.

[136] Donald Laming. Contrast sensitivity. *Vision and visual dysfunction. Macmillan Press Ltd, UK*, pages 35–43, 1991.

[137] Chang Huang, Haizhou Ai, Yuan Li, and Shihong Lao. Vector boosting for rotation invariant multi-view face detection. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 446–453. IEEE, 2005.

[138] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 654–661. IEEE, 2005.

[139] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[140] MJ McDonnell. Box-filtering techniques. *Computer Graphics and Image Processing*, 17(1):65–70, 1981.

[141] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[142] Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

[143] Muhammad Abul Hasan, Min Xu, Xiangjian He, and Ling Chen. Shot classification using domain specific features for movie management. In *DASFAA (2)*, pages 314–318, 2012.

[144] J. N. Kapur, Prasanna K. Sahoo, and A. K. C. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, pages 273–285, 1985.

[145] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. A survey of skin-color modeling and detection methods. *Pattern recognition*, 40(3):1106–1122, 2007.

[146] Shugao Ma and Weiqiang Wang. Effective camera motion analysis approach. In *Networking, Sensing and Control (ICNSC), 2010 International Conference on*, pages 111–116. IEEE, 2010.

[147] Nhat-Tan Nguyen, Denis Laurendeau, and Alexandra Branzan-Albu. A robust method for camera motion estimation in movies based on optical flow. *International journal of intelligent systems technologies and applications*, 9(3):228–238, 2010.

[148] Jurandy Almeida, Rodrigo Minetto, Tiago A Almeida, Ricardo da S Torres, and Neucimar J Leite. Robust estimation of camera motion using optical flow models. In *Advances in Visual Computing*, pages 435–446. Springer, 2009.

[149] Rodrigo Minetto, Neucimar Jerônimo Leite, and Jorge Stolfi. Reliable detection of camera motion based on weighted optical flow fitting. In *VISAPP (2)*, pages 435–440, 2007.

[150] Xinding Sun, Ajay Divakaran, and BS Manjunath. A motion activity descriptor and its extraction in compressed domain. In *Advances in Multimedia Information Processing—PCM 2001*, pages 450–457. Springer, 2001.

[151] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(7):560 –576, July 2003.

[152] Virginia Klema and Alan Laub. The singular value decomposition: Its computation and some applications. *Automatic Control, IEEE Transactions on*, 25(2):164–176, 1980.

[153] M.J. Black, Y. Yacoob, A.D. Jepson, and D.J. Fleet. Learning parameterized models of image motion. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 561 –567, jun 1997.

[154] J Friedman. Another approach to polychotomous classifcation. Technical report, Technical report, Stanford University, Department of Statistics, 1996.

[155] Sunil Lee and Chang Dong Yoo. Robust video fingerprinting for content-based video identification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(7):983–988, 2008.

[156] Alexis Joly, Olivier Buisson, and Carl Frelicot. Content-based copy retrieval using distortion-based probabilistic similarity search. *Multimedia, IEEE Transactions on*, 9(2):293–306, 2007.

[157] Sen-ching Samson Cheung and Avideh Zakhor. Efficient video similarity measurement with video signature. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13(1):59–74, 2003.

[158] A Müfit Ferman, A Murat Tekalp, and Rajiv Mehrotra. Robust color histogram descriptors for video segment retrieval and identification. *Image Processing, IEEE Transactions on*, 11(5):497–508, 2002.

[159] Mani Malek Esmaeili, Mehrdad Fatourechi, and Rabab Kreidieh Ward. A robust and fast video copy detection system using content-based fingerprinting. *Information Forensics and Security, IEEE Transactions on*, 6(1):213–226, 2011.

[160] R Cameron Harvey and Mohamed Hefeeda. Spatio-temporal video copy detection. In *Proceedings of the 3rd Multimedia Systems Conference*, pages 35–46. ACM, 2012.

[161] Changick Kim and Bhaskaran Vasudev. Spatiotemporal sequence matching for efficient video copy detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(1):127–132, 2005.

[162] Rakesh Mohan. Video sequence matching. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pages 3697–3700. IEEE, 1998.

[163] Li Chen and FWM Stentiford. Video sequence matching based on temporal ordinal measurement. *Pattern Recognition Letters*, 29(13):1824–1831, 2008.

[164] Mei-Chen Yeh and Kwang-Ting Cheng. A compact, effective descriptor for video copy detection. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 633–636. ACM, 2009.

[165] Young-tae Kim and T-S Chua. Retrieval of news video using video sequence matching. In *Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International*, pages 68–75. IEEE, 2005.

[166] Juan Manuel Barrios and Benjamin Bustos. Competitive content-based video copy detection using global descriptors. *Multimedia tools and applications*, 62(1):75–110, 2013.

[167] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[168] Travis Rose, Jonathan Fiscus, Paul Over, John Garofolo, and Martial Michel. The trecvid 2008 event detection evaluation. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–8. IEEE, 2009.

[169] Matthijs Douze, Herve Jegou, Cordelial Schmid, and Patrick Perez. Compact video description for copy detection with precise temporal alignment. *European Conference on Computer Vision*, 62(Part I):522–535, 2010.