FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

KNOWLEDGE ASSISTED DATA MANAGEMENT AND RETRIEVAL IN MULTIMEDIA DATABASE SYSTEMS

A dissertation submitted in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Min Chen

2007

To: Dean Vish Prasad
College of Engineering and Computing

This dissertation, written by Min Chen, and entitled Knowledge Assisted Data Management and Retrieval in Multimedia Database Systems, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

Yi Deng

Jainendra K. Navlakha

Nagarajan Prabakar

Mei-Ling Shyu

Keqi Zhang

Shu-Ching Chen, Major Professor

Date of Defense: March 23, 2007

The dissertation of Min Chen is approved.

Dean Vish Prasad
College of Engineering and Computing

Dean George Walker
University Graduate School

Florida International University, 2007

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude and appreciation to my dissertation advisor Professor Shu-Ching Chen for his guidance, support, suggestions and encouragement while this dissertation was being conducted. I am also indebted to Professors Yi Deng, Jainendra K Navlakha, Nagarajan Prabakar of the School of Computer Science, Professor Keqi Zhang of Department of Environmental Studies and International Hurricane Center, and Professor Mei-Ling Shyu of the Department of Electrical and Computer Engineering, University of Miami, for accepting the appointment to the dissertation committee, as well as for their suggestions and support.

The financial assistance I received from the School of Computing and Information Sciences and the Dissertation Year Fellowship from University Graduate School are gratefully acknowledged.

I would like to thank all my friends and colleagues whom I have met and known while attending Florida International University. In particular, I would like to thank Na Zhao, Kasturi Chatterjee, Khalid Saleem, Lester Melendez, Michael Armella, and Fausto Fleites for their support, encouragement, and generous help. My special thanks go to Kasturi Chatterjee and Khalid Saleem for their help with English and presentation. Finally, my utmost gratitude goes to my husband, mother, and sister, for their love, support and encouragement, which made this work possible.

ABSTRACT OF THE DISSERTATION

KNOWLEDGE ASSISTED DATA MANAGEMENT AND RETRIEVAL IN MULTIMEDIA DATABASE SYSTEMS

by

Min Chen

Florida International University, 2007

Miami, Florida

Professor Shu-Ching Chen, Major Professor

With the proliferation of multimedia data and ever-growing requests for multimedia applications, there is an increasing need for efficient and effective indexing, storage and retrieval of multimedia data, such as graphics, images, animation, video, audio and text. Due to the special characteristics of the multimedia data, the Multimedia Database management Systems (MMDBMSs) have emerged and attracted great research attention in recent years.

Though much research effort has been devoted to this area, it is still far from maturity and there exist many open issues. In this dissertation, with the focus of addressing three of the essential challenges in developing the MMDBMS, namely, semantic gap, perception subjectivity and data organization, a systematic and integrated framework is proposed with video database and image database serving as the testbed. In particular, the framework addresses these challenges separately yet coherently from three main aspects of a MMDBMS: multimedia data representation, indexing and retrieval. In terms of multimedia data representation, the key to address the semantic gap issue is to intelligently and automatically model the mid-level representation and/or semi-semantic descriptors besides the extraction of the low-level media features. The data organization challenge is mainly addressed by the aspect of media indexing where various levels of indexing are required to support the diverse query requirements. In particular, the focus of this study is to facilitate the high-level video indexing by proposing a multimodal event mining framework associated with temporal knowledge discovery approaches. With respect to the perception subjectivity issue, advanced techniques are proposed to support users' interaction and to effectively model users' perception from the feedback at both the image-level and object-level.

TABLE OF CONTENTS

CHAPTER PAGE

1 INTRODUCTION AND MOTIVATION 1
1.1 Existing Issues in MMDBMS 3
1.2 Proposed Solutions 6
1.2.1 Multimedia Data Representation 6
1.2.2 Multimedia Data Indexing 7
1.2.3 Multimedia Querying 7
1.3 Contributions 8
1.4 Scope and Limitations 11
1.5 Outline 12

2 BACKGROUND AND RELATED WORK 13
2.1 Data Management and Retrieval for Image Database 13
2.1.1 Image Data Representation 15
2.1.2 Image Feature Indexing 18
2.1.3 Content-based Image Retrieval 19
2.2 Data Management and Retrieval for Video Database 26
2.2.1 Video Shot Boundary Detection 27
2.2.2 Video Data Representation 30
2.2.3 Video Indexing and Retrieval 31

3 OVERVIEW OF THE FRAMEWORK 34
3.1 Image Database 34
3.1.1 Image Data Representation 34
3.1.2 Image Data Organization 35
3.1.3 Image Retrieval 37
3.2 Video Database 38
3.2.1 Video Data Representation 38
3.2.2 Video Indexing and Retrieval 42

4 DATA MANAGEMENT AND RETRIEVAL FOR IMAGE DATABASE 44
4.1 Image Data Representation 44
4.1.1 Global Features 45
4.1.2 Object-level Features 47
4.1.3 Semantic Network 47
4.2 Content-based Image Retrieval 53
4.2.1 Probabilistic Semantic Network-Based Image Retrieval 54
4.2.2 Hierarchical Learning Framework 63
4.2.3 Inter-database Retrieval 75

5 DATA MANAGEMENT FOR VIDEO DATABASE 84

5.1 Video Shot Detection 85

5.2 Video Data Representation 88

5.2.1 Low-level Multimodal Features 90

5.2.2 Mid-level Data Representation 95

5.3 Video Indexing and Retrieval 105

5.4 Experiments 108

5.4.1 Experimental Settings 108

5.4.2 Event Detection Performance 109

5.5 Conclusions 110

6 AUTOMATIC KNOWLEDGE DISCOVERY FOR SEMANTIC EVENT DETECTION 112

6.1 Temporal Segment Analysis for Semantic Event Detection 113

6.1.1 Temporal Pattern Analysis 114

6.1.2 Experimental Results 124

6.2 Hierarchical Temporal Association Mining 126

6.2.1 Background 127

6.2.2 Hierarchical Temporal Association Mining 130

6.2.3 Temporal rules 136

6.2.4 Experiments 138

6.2.5 Conclusions 140

7 CONCLUSIONS AND FUTURE WORK 141

LIST OF REFERENCES 147

VITA 159

LIST OF TABLES

TABLE PAGE

4.1 The relative affinity matrix $\mathcal{A}$ of the example semantic network. 48

4.2 The query access frequencies ($access_k$) and access patterns ($use_{k,m}$) of the sample images. 52

4.3 Capture R_D and R_I relationships between i_m and other images. 57

4.4 Image retrieval steps using the proposed framework. 58

4.5 The category distribution of the query image set. 59

4.6 Accuracy and efficiency comparison between Relevance Feedback method and the proposed framework. 62

5.1 Camera view descriptor. 102

5.2 Performance of corner event detection. 109

5.3 Performance of goal event detection. 110

6.1 Performance of goal event detection using temporal segment analysis. 124

6.2 Logic to find all frequent unit patterns. 133

6.3 The procedure of extended A-priori algorithm. 134

6.4 Frequent pre-actions. 135

6.5 Frequent post-actions. 135

6.6 Performance of goal event detection using temporal association mining. . . . 139

LIST OF FIGURES

FIGURE PAGE

1.1 Example images for semantic gap issue. 3

1.2 Example images for perception subjectivity issue. 4

2.1 Example results of keyword-based retrieval in Google Images. 14

2.2 Idea of the works on feature analysis and similarity measures. 20

2.3 Procedure of Relevance Feedback. 23

3.1 General video structure. 38

3.2 Overview of the proposed framework. 40

4.1 Example images 45

4.2 Probabilistic semantic network. 49

4.3 The interface of the training system. 50

4.4 Framework architecture. 55

4.5 The snapshot of a query-by-image example. 60

4.6 Performance comparison. 61

4.7 Overview of the difference between two learning schemes. (a) Idea of traditional supervised learning; (b) Idea of multiple instance learning 64

4.8 The Hierarchical Learning Framework. 67

4.9 The three-layer Feed-Forward Neural Network. 72

4.10 MMIR Experimental Results. 73

4.11 Performance comparison. 82

5.1 The multi-filtering architecture for shot detection. 85

5.2 An example segmentation mask map. (a) An example soccer video frame; (b) the segmentation mask map for (a) 87

5.3 Framework architecture. 91

5.4 Clip and frames used in feature analysis. 92

5.5 Volume of audio data. 93

5.6 (a) a sample frame from a goal shot (global view); (b) a sample frame from a close-up shot; (c) object segmentation result for (a); (d) object segmentation result for (b); (e) background variance values for frame 1 and frame 2 96

5.7 Idea of Mid-level Data Representation. 97

5.8 Three example video frames and their segmentation mask maps. 99

5.9 Example camera view. 102

5.10 Hierarchical shot view. 103

5.11 Example corner events. 104

6.1 Overview of temporal segment analysis. 114

6.2 An example two-dimensional temporal space for a time series data. 116

6.3 Overview of the algorithm for temporal segmentation. 117

6.4 Time window algorithm. 118

6.5 Time window for the Volume feature. 121

6.6 An example video sequence. 128

6.7 Hierarchical temporal mining for video event detection. 130

CHAPTER 1

Introduction and Motivation

The advances in data acquisition, generation, storage, and communication technologies have made vast amounts of multimedia data available to consumer and enterprise applications. Here, multimedia data typically refers to the digital representation of multiple media types such as graphics, image, animation, video, audio and text data. In general, the media types can be broadly classified into two categories, viz. Static and Dynamic (or time continuous) media based on whether it has time dimensions [74]. Multimedia data is blessed with a number of exciting characteristics. For instance, it can provide more effective dissemination of information in science, engineering, medicine, modern biology, and social sciences. It can also facilitates the development of new paradigms in distance learning and interactive personal/group entertainment.

Due to the proliferation of multimedia data and strong demands of multimedia applications such as TiVo, digital library, video on demand, there is a growing research interest in efficient and effective indexing, storage and retrieval of multimedia data. However, traditional Database Management Systems (DBMSs), which retrieve items based on structured data using exact matching, cannot handle multimedia data effectively because of the great differences between the characteristics of traditional textual data and multimedia data. In brief, such differences can be addressed from both the presentation and semantics perspectives.

- From a presentation viewpoint, the multimedia data is generally with huge volume and involves time-dependent characteristics that must be adhered to for coherent viewing. For example, to ensure smooth performance, videos have to be played at around 25 frames per second and a 20-minute video in MPEG format of medium frame size (320*240) with medium quality requires above 100 MB storage [74].

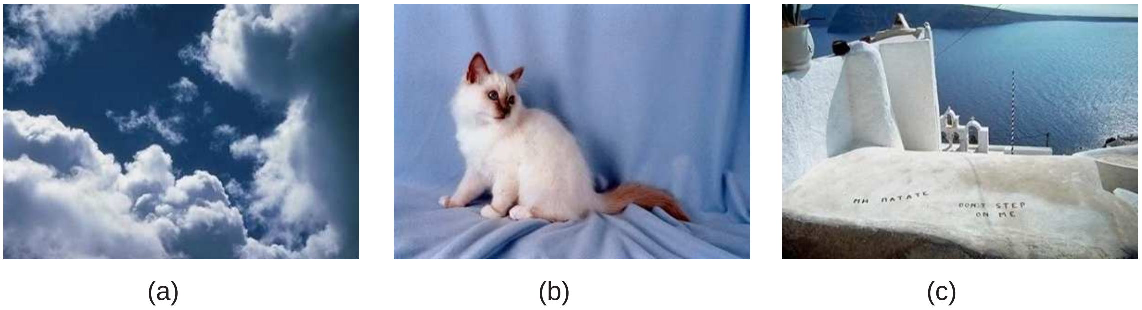
- In terms of semantics, multimedia data lacks clear semantic structure as opposed to the text data in the sense that it is represented either by a set of spatially ordered pixel values (e.g., image) or by temporally sequenced visual/audio samples (e.g., video), which inhibits the automatic content recognition by computer. In addition, multimedia data is rich in information and its meaning is sometimes fuzzy and subjective for different viewers. Therefore, multimedia data requests complex processing to derive semantics from its contents, which is not required for traditional textual data.

Due to the special characteristics of multimedia data, Multimedia Database Management Systems (MMDBMSs) have emerged and attracted great attention in recent years. In the early phase, MMDBMSs relied mainly on the operating system for storing and querying files. These were ad-hoc systems that served mostly as repositories. Later on, some MMDBMSs were proposed to handle multimedia content by providing complex object types for various kinds of media. The object-oriented style provides the facility to define new data types and operators for media such as video, image and audio. Recently, much research effort has been dedicated to capture, represent and deliver the semantic meanings of the media contents. Nevertheless, this research area is still far from maturity and there exist many open issues. Such issues include how to extract, organize, and structure features from different multimedia data for efficient retrieval, how to measure the "similarity" for different media types, how to build an easy-to-use yet sophisticated enough user interface that can construct complicated, fuzzy, and flexible queries, how to handle the spatio-temporal queries in multimedia databases, etc. As it is not possible to cover all the issues in detail within the scope of this dissertation, the primary purpose of this study is to provide a systematic and coherent solution to a number of essential issues in developing a successful MMDBMS.

The remainder of this chapter is organized as follows. In the next section, the critical research issues will be discussed. Then in Section 1.2, a brief introduction of the proposed framework will be given to address these issues. The significance and contributions of this proposed work are presented in Section 1.3. In Section 1.4, the scope and limitations of this framework are discussed. Finally, section 1.5 gives the outline of this dissertation.

1.1 Existing Issues in MMDBMS

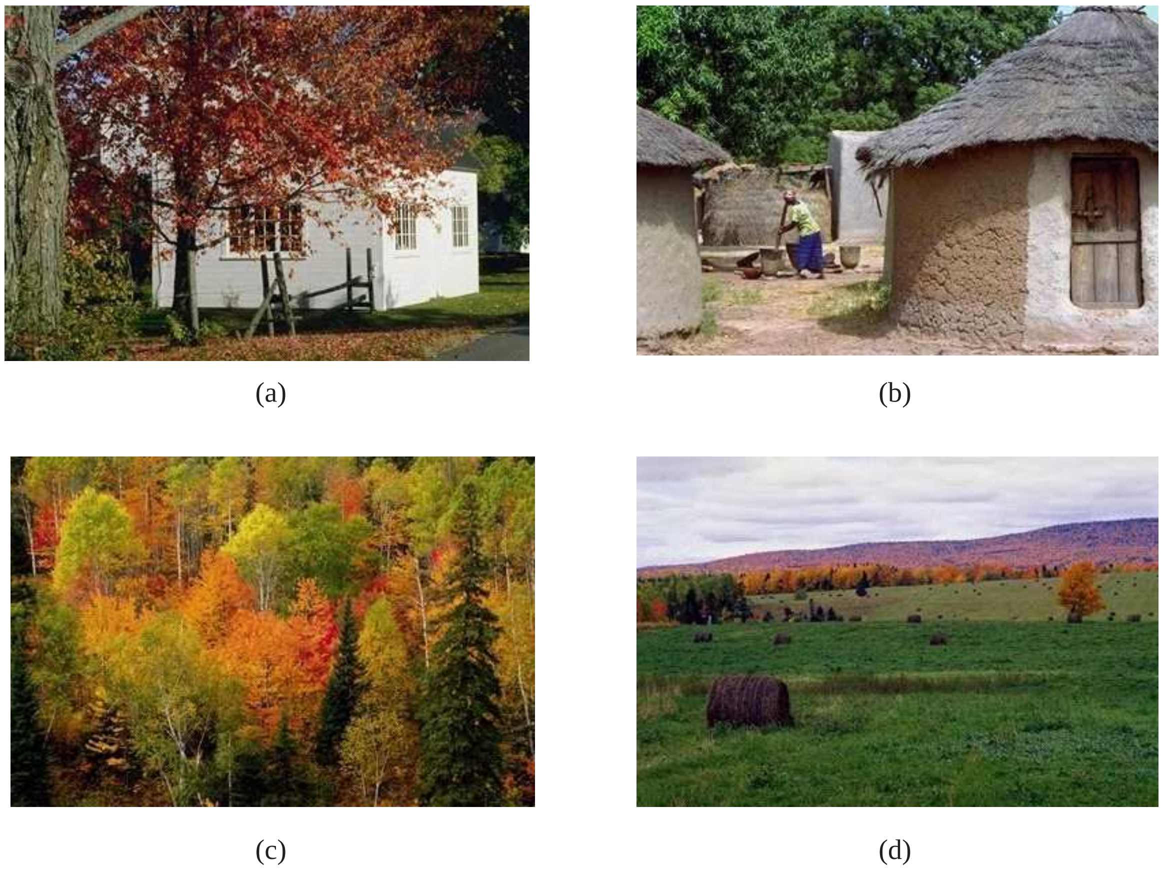
Due to the specific characteristics of the multimedia data and the emerging trend of multimedia applications, several issues are becoming especially critical for MMDBMSs.



(a) (b) (c)

Figure 1.1: Example images for semantic gap issue.

- Semantic Gap Issue. In an MMDBMS, features and attributes of media items are normally extracted, parameterized, and stored together with the items themselves. During the retrieval process, these features and attributes instead of the media items are searched and compared based on certain similarity metrics. In other words, in such a system, each media object is first mapped to a point in a certain feature space, where the features can be categorized into color [109], texture [59], shape [143], and so forth. Next, given a query in terms of a media object example, the system retrieves media objects with regard to their features [48]. However, there exists a so-called semantic gap between the representation of media and its perceived content (or semantic meaning). For instance, an image showing cloudy sky (its content) might have features of blue and white colors (its data representation) as shown in Fig. 1.1(a). If a query is issued using this image, a group of images with predominantly blue and white colors might be retrieved; among them some are actually images of “a white cat standing before a blue curtain” (Fig. 1.1(b)) or “a white castle along ocean” (Fig. 1.1(c)), etc. Note that all the images shown in this dissertation are from Corel image library unless otherwise mentioned. In real application, users typically wish to query the database based on semantics instead of low-level features. A database management system therefore requires knowledge for interpreting raw data into the content implied. Knowledge-assisted representation thus plays an essential role for multimedia database retrieval.



(a)

(b)

(c)

(d)

Figure 1.2: Example images for perception subjectivity issue.

- Perception Subjectivity Issue. The perception subjectivity problem also challenges an MMDBMS in the sense that different users might have various interpretations over a single media. In other words, in viewing the same image (e.g., Fig. 1.2(a)), different users might possess various interests in either a certain object (e.g., the house, the tree, etc.) or the entire image (e.g., a landscape during the autumn season). In this case, Fig. 1.2(b), Fig. 1.2(c), or Fig. 1.2(d), respectively, might be considered as the relevant image with regard to Fig. 1.2(a). In addition, sometimes even the same user can have different perceptions toward the same media at various situations and with different purposes. Thus, a user-friendly MMDBMS should offer a mechanism to incorporate users' feedbacks and the search engine should be equipped with an inference engine to observe and learn from user interactions.

- Data Organization Issue. Considering the huge volume of media data and high-dimensional media representations, the indexing techniques and data structures are indispensable to speeding up the search process so that the relevant media can be located quickly. Indexes of standard database systems are one-dimensional, usually hash-based or B-tree based. They are designed for standard data types, such as numbers, character strings, and in most cases are either unsuitable or insufficient for similarity matching in an MMDBMS. Over the years, many specialized indexes and data structures, such as A-tree and M-tree, have been designed for efficient multimedia retrieval based on similarity and a survey was given in [75]. However, though indexing has been studied upon the low-level features, the need for high-level indexing and retrieval is growing dramatically.

Obviously, there exist many other issues in building a full-fledged and well-performed MMDBMS. However, the above-mentioned issues are widely considered to be among the most challenging and essential problems and are the research focus of this dissertation.

1.2 Proposed Solutions

In this dissertation, a systematic and integrated framework is proposed, which addresses the above-mentioned issues separately yet coherently from the main aspects of an MMDBMS, namely multimedia data representation, indexing and retrieval. In particular, since image and video are widely deemed as the typical types in static and dynamic media categories, respectively, without loss of generality, image database and video database serve as the test beds for the proposed framework in this dissertation.

1.2.1 Multimedia Data Representation

This aspect deals with the representation of the multimedia objects to facilitate various multimedia database operations, such as media indexing, browsing, querying, and retrieval. Extensive research has been conducted to extract representative low-level features for various media, such as color [109], and texture [59] for images and visual [131], audio [133], and text [140] features for videos. Object-level features, such as object shape [143] for images or object motion [27] for videos are also studied to support operations towards the salient objects. In this dissertation, a set of low-level and object-level features that rely on MPEG-7 standard are extracted automatically from the media source. More importantly, as a key to address the semantic gap issue, mid-level and knowledge-based (high-level) data representation is explored intelligently via the knowledge discovery approaches to bridge the gaps between the semantic meaning and low-level characteristics/structure of the media.

Specifically, in terms of image database, a set of well-defined methodologies are applied to extract representative color and texture features at image level (or called global features) and object-level. A semantic network is then constructed to capture the affinity relationships among the images. Intuitively, such relationships, once captured, offer a valuable source to bridge the semantic gap since image retrieval is essentially a process to explore the relationships between the query image and the other images in the database.

Note that such techniques, though presented for image database, can be applied to video database with certain extension. In fact, in video database, it is quite common that a segment of video is represented by its key frame (a single static image which conveys the representative meaning of the video segment) and query by key frame is a basic query type in video retrieval. Therefore, effective image representation will in turn be helpful in this manner.

Compared to images, videos are more domain oriented and are rich in temporal information. Therefore, an effective mid-level representation and/or knowledge-assisted representation through temporal analysis is of great importance to address the semantic gap issue in video database.

1.2.2 Multimedia Data Indexing

Data indexing is an essential process for effective data organization and fast data searching. Structured organization of video data is especially critical due to its huge data volume and complicated spatial-temporal characteristics. Video indexing is thus the basis to build large video archives which allow efficient retrieval, browsing and manipulation. Conventionally, video data have to be manually annotated with keywords. However, it is time consuming and most likely incomplete. Therefore, the focus of this dissertation is to develop automatic video analysis and indexing techniques. In particular, an effective data analysis and classification framework is proposed to capture the important and interesting activities (called events, such as goal events, traffic accidents) and high-level semantic features (called concepts, such as commercial, sports) in the video data, which in turn lead to the high-level video indexing.

1.2.3 Multimedia Querying

Querying in a multimedia database normally differs greatly from that in a traditional database in the sense that the queries can contain multimedia object issued by the user (called Query-By-Example or QBE) and the query results can be based not on perfect matches but on degrees of similarity. Such a process is called Content-Based Retrieval (CBR). Compared to image queries, users' perceptions towards a certain video segment are generally more consistent and well-defined with the assistance of its context. Therefore, in this dissertation, the perception subjectivity issue is systematically studied for image retrieval, where both the general concepts and individual user's specific interests are taken into consideration.

Specifically, a long-term learning (or called log-based retrieval) approach is devised to capture users' general concepts by stochastically analyzing the historical feedback logs accumulated in the database. As will be discussed in Section 2.1.3, this mechanism targets reducing the overhead incurred during on-line users' relevance feedback and addresses the "cold-start" problem in long term (collective) learning. Meanwhile, to acknowledge individual user's specific query interests, a real-time query refinement scheme is proposed, which is conducted through user interaction and the similarity metrics is re-visited to take into consideration the user perception. This scheme is integrated seamlessly with the long-term learning process in the proposed framework, which can be stated from two perspectives. First, long-term learning provides an effective mechanism to speed up the convergence of the real-time query refinement process. Second, the relevance feedback information is accumulated in the database and the long-term learning is triggered periodically when the number of accumulated feedbacks reaches a certain threshold.

In summary, the main objective of this dissertation is to explain the working principles of a novel data management and retrieval system whose main functionality is to endow its users with an easy-to-use, effective, and efficient scheme for retrieving the required multimedia information.

1.3 Contributions

In this dissertation, a new paradigm is presented for effective data management and retrieval of multimedia data. The proposed system tries to achieve its objectives by developing novel and effective approaches to tackle the issues addressed in Section 1.1. The major contributions are listed as follows.

1. Different from the common approaches which try to capture the semantic content of an individual image (it is more difficult and most likely incomplete), a probabilistic semantic network is constructed in this study to represent the semantic relationships among images. Such a network is useful because image retrieval is actually a process to explore the relationships between the query image and the other images in the database. In addition, in contrast to the work proposed in [76] that requires extra manual effort in labeling the images, this framework provides the capability to accumulate the previous feedback information and automatically mine the semantic relationships among the images to construct and update the probabilistic semantic network. Therefore, instead of starting each query with the low-level features, the semantic information is gradually embedded into the framework to improve the initial query results.

2. Besides using accumulative learning to bridge semantic gaps in image data representation, the proposed framework also supports the query refinement to accommodate individual user's query interests in real-time. In particular, a temporary semantic subnetwork is extracted from the semantic network and updated based on the current user's interests. For the sake of system efficiency and avoiding the bias caused by a single user, such update is conducted only on the temporary semantic subnetwork, without affecting the original semantic network. In the meanwhile, the individual user's feedback is collected continuously, and the update of the whole semantic network is triggered only when the number of accumulated feedbacks reaches a threshold. Such update is conducted off-line to enable accumulative learning while maintaining efficiency.

3. As an effort to further extend the semantic network based data representation and retrieval scheme, a unified framework incorporating Multiple Instance Learning technique is developed to explore the high-level semantic concepts in a query from both the object-level and image-level and to address the needs of serving the specific user's query interest as well as reducing the convergence cycles.

4. In terms of video database, a systematic framework is devised for video content analysis and indexing, which consists of three primary processes: structure analysis, multimodal data representation, and abstraction (i.e., high-level indexing). A set of novel techniques and methodologies are applied in each component and integrated seamlessly to both reduce the processing time and improve the system accuracy. In particular, to effectively link the low-level features to the content and structure of video data, a group of mid-level descriptors are introduced, which are deduced from low-level feature representations and are motivated by high-level inference. Such mid-level descriptors offer a reasonable tradeoff between the computational requirements and the resulting semantics. In addition, the introduction of mid-level descriptors allow the separation of domain specific knowledge and rules from the extraction of low-level features and offers robust and reusable representations for high-level semantic analysis using customized solutions.

5. With the ultimate goal of developing an extensible video content analysis and indexing framework that can be robustly transferred to a variety of applications, a critical aspect is to relax the need for the domain knowledge, and hence to reduce the manual efforts in selecting the representative patterns and defining the corresponding thresholds. For this purpose, a novel temporal segment analysis approach and temporal association rule mining scheme are proposed, which are motivated by the fact that temporal information of a video sequence plays an important role in conveying the video content. The advantage of such approaches is that they largely

improve the extensibility and flexibility of the proposed video content analysis and indexing framework.

1.4 Scope and Limitations

The proposed framework has the following assumptions and limitations.

1. In the proposed image data representation and retrieval approaches, various assumptions are made in terms of the amount of noisy data contained in the image database. For instance, it is presumed that the image quality is reasonably good. In addition, for the proposed long-term learning framework, it is assumed that while a certain user might introduce the noise information into the query log, the rate is negligibly low. Some of the assumptions might not hold in real-world applications, especially in this era of information explosion. Therefore, the construction of the noise-tolerate mechanism may be required, where the techniques like outlier detection, fuzzy logic, etc. can be introduced for this purpose.

2. Though a set of advanced techniques, such as semantic network based data representation and retrieval, temporal segment analysis and temporal association mining, etc., are proposed in this framework to effectively alleviate the semantic gap and perception subjectivity issues, it is very challenging to solve these issues completely and ultimately. In fact, it is a general agreement in the multimedia database research society that it is difficult to build a fully-automatic, general-purpose MMDBMS which can understand media content and match users' perception perfectly.

3. Object segmentation remains an open issue and the results are far from satisfactory according to the current state of the art in computer vision, pattern recognition and image processing. Intuitively, one of the reasons lies in the fact that the analysis of low-level features alone cannot provide accurate descriptions for semantic objects. As a result, sometimes an object is segmented into several regions (called

over-segmentation) or multiple objects are merged into one segment (called under-segmentation). In this framework, though the dependency on the segmentation results has been largely relaxed, the performance of several components such as region-based retrieval and shot-boundary detection can be potentially affected.

1.5 Outline

The organization of this dissertation is as follows. In Chapter 2, the literature reviews are given in the areas of content-based indexing and retrieval for image and video data, with the focus on the existing approaches in addressing the semantic gap, perception subjectivity and data management challenges.

Chapter 3 describes the proposed multimedia data management and retrieval framework for the multimedia database systems. Each component of the framework will be discussed in detail.

The current stand of the proposed data management and retrieval for image database is presented in Chapter 4, where the focus is on the automatic knowledge discovery for image representation and retrieval.

In Chapter 5, data management and indexing for video database is discussed. Specifically, a mid-level data representation and high-level event detection framework is detailed.

In Chapter 6, to further assist the video high-level indexing, knowledge discovery approaches are discussed to intelligently explore the characteristic temporal patterns for event detection.

In Chapter 7, the conclusions are given with the proposed future work.

CHAPTER 2

Background and Related Work

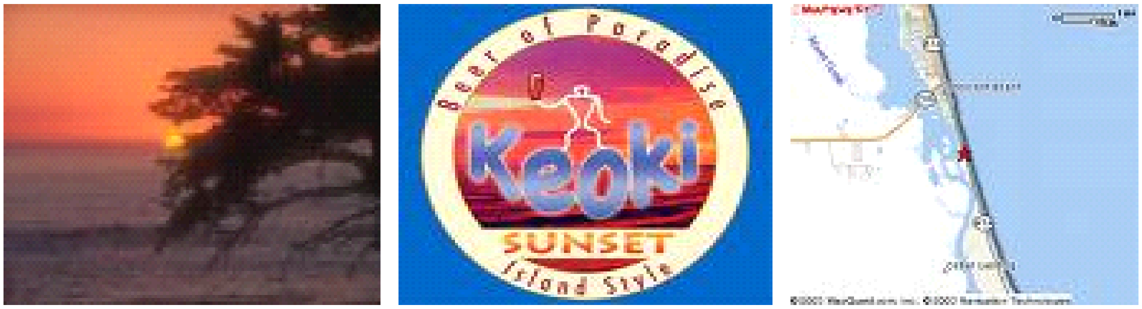
As discussed in Section 1.2, image database and video database are selected as the test bed for the proposed framework in this dissertation. In this chapter, the existing approaches and methodologies of data management and retrieval for image and video database systems are summarized.

2.1 Data Management and Retrieval for Image Database

Digital images hold an important position among all the multimedia data types. They are central to a wide variety of applications, ranging from medicine to remote sensing, and play a valuable role in numerous human activities, such as entertainment, law enforcement, etc. In the literature, there are three main approaches to support image retrieval [74], which in turn affects the design of image databases.

- In the first type, image contents are modeled as a set of predefined attributes (such as image category, subject, etc.) extracted manually and managed within the framework of conventional database management systems. Queries are specified by using these attributes. Thus images can be indexed and retrieved by using a powerful relational database model [85]. Obviously, the major drawback of this approach is that these attributes may not be able to describe the image contents completely, and the types of queries are limited to those based on these attributes.

- The second approach uses textual descriptions (keywords) to describe (annotate) images and employs Information Retrieval (IR) techniques to carry out image retrieval. Text can describe the high-level abstraction contained in images. However, it has two major issues. First, image annotation requires a prohibitive amount of labor when the size of image database becomes large. The other, and probably most essential, drawback results from the difficulty of capturing the rich image con-



(a) (b) (c)

Figure 2.1: Example results of keyword-based retrieval in Google Images.

tents using a small number of keywords and the subjectivity of human perception involved in the annotation process [69]. As an example, Google Images [39] currently supports keyword-based retrieval scheme. Given a query keyword "Sunset" with the intention of retrieving sunset landscape images, the retrieval results might not be satisfactory, as illustrated in Fig. 2.1. As can be seen, though Fig. 2.1(a) shows a sunset scene, Fig. 2.1(b) (Keoki *Sunset* Bottled) and Fig. 2.1(c) (Driving directions to *Sunset* beach) are also returned as they both were annotated with the keyword of "Sunset."

- The third approach uses global or object-level image features to index and retrieve images. This approach is generally called Content-Based Image Retrieval (CBIR) as the retrieval is based on pictorial contents. The advantage of this approach is that the indexing and retrieval process can be automatically performed and conveniently implemented. However, it suffers from the semantic gap and perception subjectivity issues as discussed in Section 1.1.

Currently, the research and design of image database management systems aim to support the third approach. Therefore, a literature review of the image data representation, indexing and retrieval aspects in such systems is given in the following sections.

2.1.1 Image Data Representation

Feature extraction is the basis for an image database system and constitutes the first stage of indexing images by content. Features can be categorized as general or domain-specific [69]. General features typically include color, texture, shape and sketch, whereas domain-specific features are applicable in specialized domains such as human face recognition or fingerprint recognition. As the target of this dissertation is towards general image databases, in this section only the general features are introduced. Note that each feature may have several representations. For example, as will be discussed below, color histogram color moments, and color sets are representations of the image color feature.

Color Features

Color is one of the most recognizable elements of image content and is widely used as image data representation because of its invariance with respect to image scaling, translation, and rotation. The key issues in color feature extraction include the color space and color quantization.

- Color Space. The commonly used color spaces include RGB, HSL, and CIELAB. Here RGB stands for Red-Green-Blue which are primary colors used to compose any other colors. RGB is device-dependent and normally used on monitors. HSL denotes Hue, Saturation and Luminosity. Hue is the perception of the nuance. It is the perception of what one sees in a rainbow. The perception of Saturation is the vividness and purity of a color. For example, a sky blue has different saturation from a deep blue. Luminosity, also called brightness, is the perception of an area to exhibit more or less light. Although the representation of the colors in the RGB space is quite adapted for monitors, HSV space is preferred for a human being. In terms of CIELAB, the CIE defined the Lab spaces in order to get more uniform and accurate color models, where L defines lightness, a denotes red/green value, and b indicates the yellow/blue value. It is worth mentioning that MPEG-7 XM V2 supports RGB and HSV color spaces, and some linear transformation matrices with reference to RGB [82].

- Color Quantization. Color quantization is used to reduce the color resolution of an image. Using a quantized color map can considerably decrease the computational complexity during image retrieval. In MPEG-7 XM V2, three quantization types are supported: linear, nonlinear, and lookup table [82].

The commonly used color feature representations in image retrieval include color histogram, color moments, and color sets.

Texture Features

Texture refers to visual patterns with properties of homogeneity that do not result from the presence of only a single color or intensity [107]. Although the ability to retrieve images on the basis of texture similarity may not seem very useful, the ability to match on texture similarity can often be useful in distinguishing between areas of images with similar color. Typical textural features include contrast, uniformity, coarseness, roughness, frequency, density, and directionality [72]. Texture features usually contain important information about the structural arrangement of surfaces and their relationship to the surrounding environment. There are two basic classes of texture descriptors: statistical model-based and transform-based. The former explores the gray-level spatial dependence of textures and then extracts meaningful statistics as texture representation, which were adopted in some well-known CBIR systems such as QBIC [35] and MARS [86]. As for transform-based texture extractions, some commonly used transforms are the discrete cosine transform (DCT transform), the Fourier-Mellin transform, the Polar Fourier transform, and the wavelet transform.

Shape Features

Unlike texture, shape is a fairly well defined concept, and there is considerable evidence that natural objects are primarily recognized by their shape. To extract shape features, two steps are involved, namely, object segmentation and shape representation.

- Object Segmentation. The existing segmentation techniques include the global threshold-based, the region-growing, the split-and-merge, the edge-detection-based, the texture-based, the color-based, and the model-based techniques [69]. Generally speaking, it is difficult to achieve a precise segmentation owing to the complexity of the individual object shape, the existence of shadows and noise, etc.

- Shape Representation. Once objects are segmented, their shape features can be represented and indexed. In general, shape representations can be classified into three categories: boundary-based representations (e.g., Fourier descriptor), region-based representations (e.g., moment invariants) and combined representations (i.e., the integration of several basic representations such as moment invariants with Fourier descriptor).

Other features

In the literature, many other types of image features have also been proposed, which mainly rely on complex transformations of pixel intensities to yield better image representations with regard to human descriptions. Among them, the most well-known technique is to use the wavelet transform to model an image at several different resolutions.

Studies show that the results are often unsatisfactory in real applications with the use of a single class of descriptors. A strategy to potentially improve image retrieval is to combine multiple heterogeneous features, which result in multidimensional feature vectors. In an image database, such vectors are often used to measure the similarity of two images by calculating a descriptor distance in the feature space. For a large-scale image database, a sequential linear search fails to provide reasonable efficiency. Thus feature indexing becomes necessary.

2.1.2 Image Feature Indexing

In traditional DBMSs, data are indexed by key entities, where the most popular indexing techniques are B-tree and its variations. In an image database, the images should be indexed based on extracted inherent visual features such as color and texture to support an efficient search based on image contents. As mentioned above, an image can be represented by a multidimensional feature vector, which acts as the *signature* of the image. Intuitively, this feature vector can be assumed to be associated with a point in a multidimensional space. For instance, assume images in an image database are represented by N-dimensional feature vectors. Retrieving similar images to a query image then is converted to the issue of finding the indices of those images in the N-dimensional search space whose feature vectors are within some threshold of proximity to the point of the query image. This indexing structure is widely known as Multidimensional Access Structure (MAS).

The B-tree related indexing techniques are not suitable to index the high-dimensional features. Thus in the literature, a number of multidimensional indexing techniques have been proposed. For instance, in [24], M-tree was proposed to organize and search large data sets in metric spaces. [96] proposed an index structure called A-tree (Approximation tree) for similarity search of high-dimensional data using relative approximation. A KVA-File (kernel VA-File) that extends VA-File to kernel-based retrieval methods was proposed in [49]. A survey of the techniques and data structures, such as k-d tree, quad-tree, R-tree and its variants (R^+ tree and R^* tree), VAM k-d tree, VAMSplit R-tree, and 2D h-tree, for efficient multimedia retrieval based on similarity was given in [75]. Among them, the R-tree and its variants are the most popular. In brief, an R-tree is a B-tree-like indexing structure where each internal node represents a k-dimensional hyper-rectangle.

Experiments indicate that though R-trees and R^* trees work well for similarity retrieval when the dimension of the indexing key is less than 20, the performance of these tree-structured indices degrades rapidly for a higher dimensional space [69]. Therefore, dimension reduction might be required before employing the multidimensional indexing technique upon the feature vectors. Karhunen-Loeve Transform (KLT) and its variations have been widely used in dimension reduction in many areas such as features for facial recognition, eigen-images and principal component analysis [58].

2.1.3 Content-based Image Retrieval

Multimedia information, typically image information, is growing rapidly across the Internet and elsewhere. With the explosive growth in the amount and complexity of image data, there is an increasing need to search and retrieve images efficiently and accurately from image databases. However, as discussed earlier, the traditional query-by-keyword is not suitable for image retrieval for the following reasons: 1) Keyword-based annotation is extremely labor-intensive in processing the voluminous images; and 2) Due to the rich semantics of the images and the subjectivity of human perception, it is difficult to choose the proper keywords for the images. To solve these problems, instead of indexing and retrieving by keywords, Content-Based Image Retrieval (CBIR) was proposed to retrieve images based on their content.

The existing works in CBIR can be roughly classified into the following four categories:

1. Feature Analysis and Similarity Measures

   Many early-year studies on CBIR focused primarily on feature analysis and similarity measures [89][146]. Fig. 2.2 illustrates the basic idea of these approaches. Given the images in the image database as shown in Fig. 2.2(a), their features of all the images will be extracted. Now if the user wants to retrieve the images similar to this query image as highlighted by the green rectangle in Fig. 2.2(a), the similarity values between the query image and other images will be calculated in the image

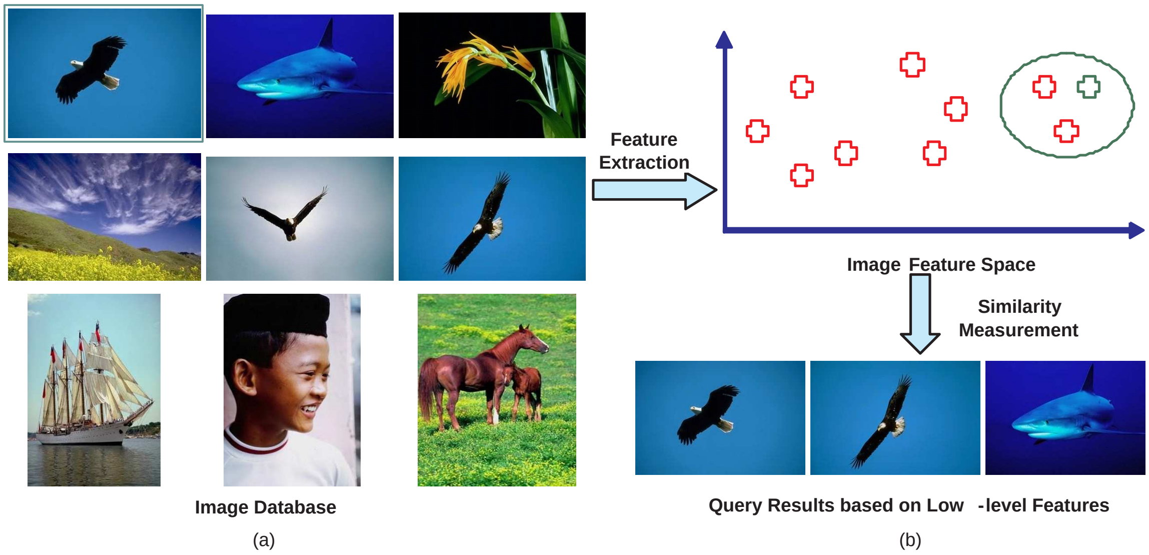


Figure 2.2: Idea of the works on feature analysis and similarity measures.

feature space and the most similar images are returned to the user as shown in Fig. 2.2(b). Some research prototypes and commercial systems have been implemented for CBIR. In the Virage system [40], image content is given primarily in terms of the properties of color and texture. The QBIC system of IBM [35] provides the support for queries on color, texture and shape. The photobook [88] system supports queries by image content in conjunction with text queries. However, due to the semantic gap and the perception subjectivity issues, it is extremely difficult to discriminate the images by solely relying on the similarity measure upon the low-level features in the real-world image databases [50]. As shown in Fig. 2.2(b), the retrieved images might include the misidentified "fish" image and omit the correct one (e.g., the "eagle" image centered in Fig. 2.2(a)).

2. Relevance Feedback (RF)

   A variety of RF mechanisms from heuristic techniques to sophisticated learning techniques have been proposed and actively studied in recent years to mitigate the semantic gap by modeling the user's subjective perception from the user's feedback [95][118]. The principle of RF is to adjust the subsequent queries by altering the position of the query point (or called the query center) and/or the feature weights based on the information gathered from the user's feedback, which can be regarded as a form of supervised learning. As illustrated in Fig. 2.3, the basic idea is as follows: Once a query is formulated, the system returns an initial set of results. In this process, all features used in the similarity metric are considered equally important because the user's preference is not specified. Using the initial result set, the user can give some positive or negative feedback to the system. For example, labels such as "relevant" and "irrelevant" can be attached to the images. The query, augmented by labeled images, is then resubmitted and processed by the search engine. The system will thereafter refine the query and retrieve a new list of images. Hence, the key issue in RF is how to incorporate positive and negative examples in query and/or in the similarity refinement. There are two main approaches called query point movement (query refinement) and re-weighting (similarity measure refinement).

- The query point movement method essentially tries to improve the estimate of the "ideal query point" by moving it towards good example points and away from bad example points. The frequently used technique to iteratively improve this estimation is Rocchio's formula.
- The central idea behind the re-weighting method is to re-weight different features during the search, reflecting the importance attached to them by the user. It enhances the importance of those dimensions of a feature vector that help in retrieving the relevant images and reduce the effects of those that hinder this process. Typically, the empirical standard deviation of each feature is computed from the example set and its inverse is used as weight. That is, if the variance of the good examples is high along a principle axis j (i.e., j^{th}

feature), it can be deduced that the values on this axis are not very relevant to the input query and the importance of this feature is relatively low. Therefore, a low weight w_j is assigned on it.

From past research studies, RF has been shown as an effective scheme to improve the retrieval performance of CBIR and has already been incorporated as a key part in designing a CBIR system. Examples of such systems include MARS [86], IRIS [136], WebSEEK [108], PicHunter [26], etc. However, those RF-based systems have two major limitations as follows.

- RF estimates the ideal query parameters only from the low-level image features. Due to the limited power of the low-level features in representing the high-level semantics, it is quite common that the relevant samples are scarce in the initial query or the relevant images are widely scattered in the feature space. As a result, RF technique is often inadequate in learning the concepts [55]. For instance, it typically takes quite a number of iterations to achieve convergence of the learning process to obtain the high-level concepts. In many cases, the desired query results could not be achieved even after a large number of user interactions.
- Though the feedback information provided in each interaction contains certain high-level concepts, it is solely used to improve the current query results for a specific user. In other words, no mechanism is included in these systems to memorize or to accumulate the relevance feedback information to improve both the current query accuracy and the future system performance.

Furthermore, most of the existing RF based applications regard each image as a whole, which often fails to produce satisfactory results when the user's query interest is just the salient region(s) in the image.

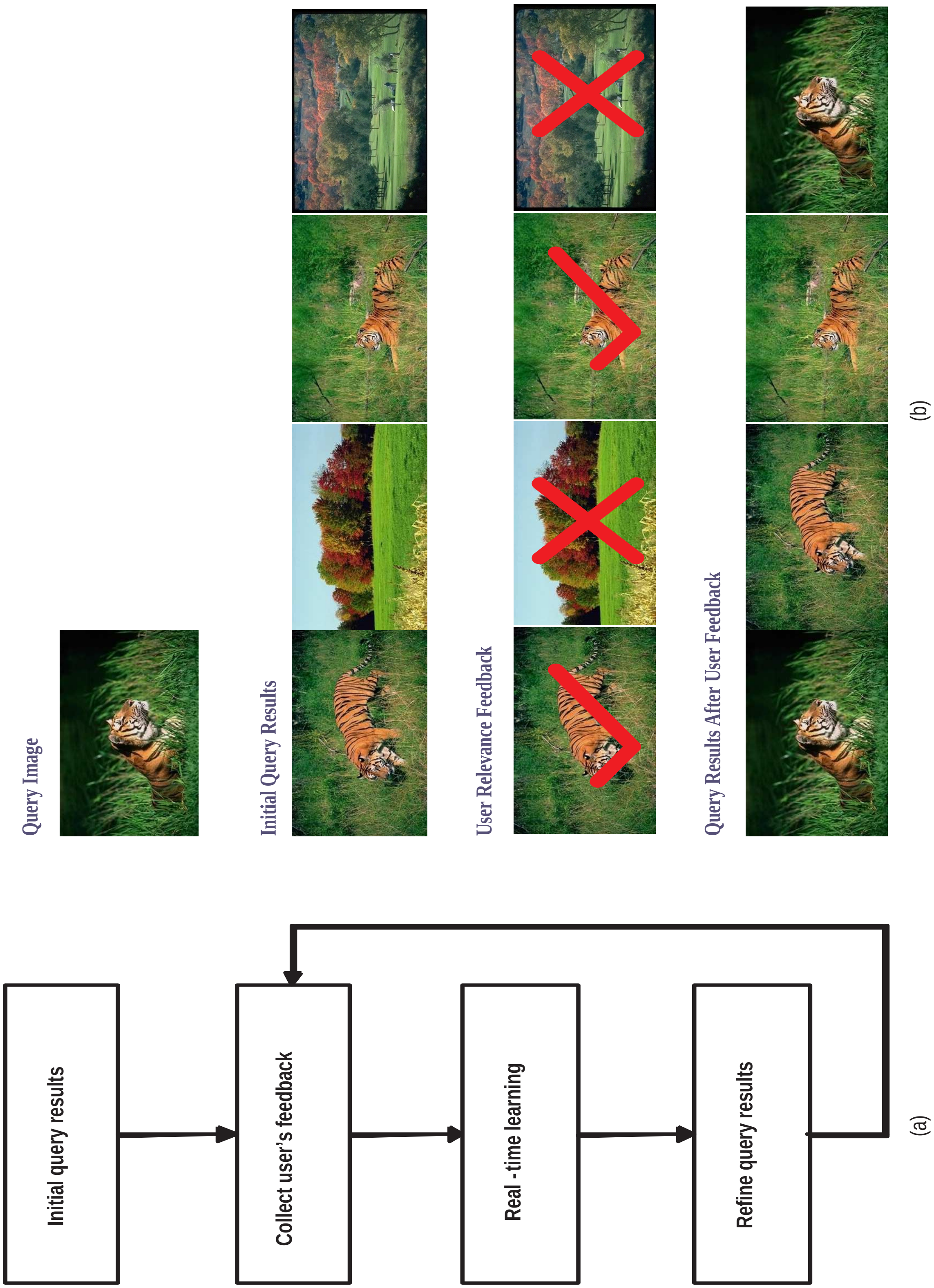


Figure 2.3: Procedure of Relevance Feedback.

3. Region-based approaches

   With the assumption that human discernment of certain visual contents could be potentially associated with the semantically meaningful object(s) in the image, region-based retrieval [21][56] and MIL [13] techniques offer an alternative solution by decomposing the images into a set of homogeneous regions for analysis. Each region roughly corresponds to an object and is represented by a set of local image features. The similarity measurements are then applied at the object/region level. As a result of continuous effort towards this area, some region-based image retrieval systems have been proposed. For example, Blobworld [4] is an early region-based image retrieval system that segments the images into blobs based on color and texture features, and queries the blobs by using some high-dimensional index structure. For each image, a similarity score is given by a fuzzy combination over the similarity scores between the query blobs and their most similar blob in that image. However, the multi-region queries remain unclear and unaddressed in this work. The SIMPLicity [124] system uses the integrated region matching technique (IRM) to allow many-to-many matching between regions in two images. WALRUS [84] is another region-based retrieval system, segmenting images by using wavelets. The use of wavelets for segmentation has been promising. In its retrieval process, the sum of the sizes of all the retrieved regions for each image is calculated, and only those images with their matched region sizes exceeding some threshold are returned. In [56], an indexing schema customized especially for region-based image retrieval was proposed. Other systems in this category include [1] [20][70]. It is worth noting that, as will be discussed in Section 4.2.2, to some extent the MIL technique might be considered as a hybrid of the RF technique and the region-based approach. However, semantically accurate image segmentation is an ambitious long-term goal for computer vision researchers, which highly limits the performance of these approaches. Here, semantically accurate image segmentation means the capability of building a one-to-one mapping between the segmented regions and the objects in the image [21]. In addition, the assumption of the existence of salient object(s) in the images does not always hold.

4. Log-based retrieval (or called long-term learning)

   Due to the complexity of image understanding, the regular learning techniques, such as RF and MIL, need quite a number of rounds of feedback to reach satisfactory results. In addition, all the feedback obtained during the RF or MIL process is solely used to improve current query results for a specific user. In other words, no mechanism is provided to memorize or to accumulate the valuable relevance feedback information to improve both the current query accuracy and the future system performance. Consequently, log-based retrieval was proposed recently [50][106], which seeks to speed up the convergence process to capture the high-level semantic concepts in a query with the assistance of the historical feedback logs accumulated in the database system from the long-term learning perspective. However, most of the existing log-based retrieval frameworks solely capture the general user concepts but fail to adjust or customize the high-level semantic concepts in a query with regard to a specific user. Also, similar to most of the RF techniques, they have difficulty in propagating the feedback information across the query sessions toward the region or object level.

As can be seen, by acting alone the above-mentioned approaches have certain limitations in terms of retrieval accuracy and/or processing costs. Therefore, a few efforts have been directed to propose the integrated frameworks to improve the retrieval performance. In [47], the authors suggested to incorporate the RF technique with the Singular Value Decomposition (SVD) based long term learning. In addition, Hoi et al. [50] studied the log-based relevance feedback for the purpose of improving the retrieval performance and reducing the semantic gap in CBIR. However, these approaches solely direct the focus on the image level. In our recent work [22], we extended our research efforts to the object level by incorporating the Latent Semantic Indexing (LSI) based long-term learning and One-class Support Vector Machine (SVM) based MIL technique. However, to record the query logs, the users are asked to pick the region of interest in the segmented image, which imposes a heavy burden on the users.

2.2 Data Management and Retrieval for Video Database

An enormous amount of video data is being generated these days all over the world. This requires efficient and effective mechanisms to store, access, and retrieve these data. Video streams are considered the most complex form of multimedia data because they contain almost all other forms such as images and audio in addition to their inherent temporal dimension. One promising solution that enables searching multimedia data, in general, and video data in particular is the concept of content-based search and retrieval. Similar to image database, data representation, indexing and retrieval need to be addressed for the video database systems. However, different from static images, a video sequence consists of a sequence of images taken at a certain rate. Therefore, if these frames are treated individually, indexing and retrieval might not be efficient. Consequently, most of the proposed video indexing and retrieval prototypes are constructed with the following two major phases [33]:

1. Database Population Phase

   Normally, this phase consists of the following steps.

   - Shot Boundary Detection. Since video is normally made of a number of logical units or segments, the purpose of this step is to partition a video stream into a set of meaningful and manageable segments, which then serve as the basic units for indexing. Shot-based video indexing and retrieval is the approach generally adopted in the video database, where a shot is defined as a short sequence of contiguous frames that signify a single camera operation.

- Key Frames Selection. This steps attempts to summarize the information in each shot by selecting representative frames that capture the salient characteristics of that shot.
- Extracting Low-Level Features from Key Frames. During this step, a number of low-level spatial features (color, texture, etc.) are extracted in order to use them as indices to key frames and hence to shots. Temporal features (e.g., object motion) can be used too.

2. Retrieval Phase

   In this stage, a query is presented to the system that in turn performs similarity matching operations and returns similar data (if found) back to the user. One technique that is commonly used to present queries to video databases is so-called Query By Example (QBE). In this technique, an image or a video clip is presented to the system and the user requests the system to retrieve similar items.

Consequently, the related works in video data management and information retrieval are discussed in the following sections.

2.2.1 Video Shot Boundary Detection

The first step is to segment the video into shots, a step commonly called video temporal segmentation, partition, or shot boundary detection. A camera break (or called shot cut) is the simplest transition between two shots, where consecutive frames on either side of a camera break display a significant quantitative change in content. More sophisticated camera operations include dissolve, wipe, fade-in, and fade-out. Here, fade-in means a scene gradually appears. Fade-out is when a scene gradually disappears. Dissolve is when one scene gradually disappears while another gradually appears. Wipe is when one scene gradually enters across the frame while another gradually leaves. Such special effects involve much more gradual changes between consecutive frames than does a camera break and require a more sophisticated approach.

The key issue in shot detection is how to measure the frame-to-frame differences. A number of difference measures between frames have been proposed. The most simple measure is the sum of pixel-to-pixel differences between neighboring frames [142]. If the sum is larger than a preset threshold, a shot boundary is said to exist between these two frames. This method is not effective and many false shot detections will be reported because it is sensitive to object and camera movement.

The second method measures color histogram distance between neighboring frames [114]. The principle behind this method is that object/camera motion causes little histogram difference. Thus if a large difference is found, it is highly possible that a camera break occurred.

The above shot detection techniques rely on a single frame-to-frame difference threshold for shot detection. However, they have difficulty in detecting shot boundaries when the change between frames is gradual. In addition, these techniques do not consider spatial color distribution. Different techniques are needed to tackle these problems.

The difference values within a fade-in, fade-out, dissolve, and wipe operation tend to be higher than those within a shot but significantly lower than the shot cut threshold. Intuitively, a single threshold might not work, since in order to capture gradual transition, the threshold must be lowered significantly, which results in many false detections. To solve this problem, Zhang et al. [142] developed a twin-comparison technique to detect both normal camera breaks and gradual transitions. In general, it is hard to correctly determine gradual transitions. Trying to improve the success rate, [138] proposed a shot detection technique based on wavelet transformation. Their technique is based on the assumption that during fade-in, fade-out, and dissolve, the high frequency component of the image is reduced. As discussed in [74], ideally the frame-to-frame distances used for shot detection should have a distribution close to zero with very little variation within a shot and significantly larger than those between shots. However, the frame-to-frame distances of common videos do not have this type of distribution due to object and camera motion and other changes between frames. To improve the shot detection performance, a filter was proposed in [87] to remove the effects of object and camera motion so that the distribution of the frame-to-frame distance is close to the ideal distribution. Besides the works based on the color or intensity histograms, [149] proposed a shot detection method based on edge detection. In addition, Haas et al. [41] presented a method of using the motion within the video to determine the shot boundary locations.

There are also some works which carry out video segmentation and indexing directly based on compressed data [52][66]. Two main types of information used are Discrete Cosine Transform (DCT) coefficients and motion information. In MPEG 1 and MPEG 2, DCT is applied to each I block and differential block. Among the 64 DCT coefficients of each block, the first coefficient, called the Direct Current (DC) coefficient, represents the average intensity of that block. The DC image is 64 times smaller than the original image, but contains the main features of the original image. Therefore, many researchers have proposed to perform video segmentation based on DC images where the frame-to-frame distance measures can be used with minor update [66]. Another type of information that is used for video segmentation is motion information. In brief, the directional information of motion vectors are used to determine camera operations such as panning and zooming. Then the number of bidirectionally coded macroblocks in B frames is used for shot detection. If a B frame is in the same shot as its previous and next reference pictures, most macroblocks can be coded by using bidirectional coding. Therefore, if the number of bidirectional coded macroblocks is below a certain threshold, it is likely that a shot boundary occurs around the B frame [129].

After the video segmentation process, the next step is to represent and index each shot so that shots can be located and retrieved quickly in response to queries.

2.2.2 Video Data Representation

The most common way is to represent each shot with one or more key frames or representative frames. The reference frame(s) captures the main contents of the shot, whose features such as color and shape, as discussed in Section 2.1.1 can be extracted. An important issue in video data representation is actually related to the problem of how to choose the representative frame(s). It can in turn be decomposed into two subproblems as 1) how many reference frame(s) should be used in a shot, and 2) how to select the corresponding number of reference frame(s) within a shot.

A number of methods have been proposed to address these issues as follows.

- The first method uses one reference frame per shot where the first frame is normally picked. The limitation of this method is that it does not consider the length and content changes of shots [81].

- The second method assigns the number of reference frame(s) to shots according to their length. In this method, a segment is defined as a video portion with a duration of one second. Then an average frame is defined so that each pixel in this frame is the average of pixel values at the same grid point in all frames of the segment. Finally the frame within this segment that is most similar to the average frame is selected as the representative frame. However, this approach solely considers the shot duration and ignores its content.

- In the third method, each shot is divided into subshots which are detected based on changes in content with respect to motion vectors, optical flow, etc. Then the histograms of all the frames in the subshot are averaged and the frame whose histogram is the closest to this average histogram is selected as the reference frame.

In general, the choice of reference frame selection method is application dependent. Besides the features extracted at image-level or object-level from the reference frames, shot-level motion information is also generally extracted from optical flow or motion vectors, to capture the temporal or motion information contained in the video [111].

2.2.3 Video Indexing and Retrieval

Indexing video data is essential for providing content based access. Since the indexing effort is directly proportional to the granularity of video access (or retrieval interests), in this dissertation, video indexing and retrieval are discussed and considered as a concrete unit. In the literature, video indexing and retrieval can be broadly classified into four main categories [3].

- High-level Indexing and Retrieval. This approach uses a set of predefined index terms for annotating video. The index terms are organized based on a high level ontological categories like action, time, and space. In this case, video can be indexed and retrieved based on annotation using the traditional IR techniques. The high level indexing techniques are primarily designed from the perspective of manual indexing or annotation and require considerable manual efforts. However, it is still widely used because automatic high-level video content understanding is currently not feasible for general video. Alternatively, many videos have associated transcripts and subtitles that can be directly used for video indexing and retrieval. Thirdly, if subtitles are not available, speech recognition can be applied to the sound track to extract spoken words, which can then be used for indexing and retrieval. However, this approach is still very challenging because the performance of speech recognition is still far from satisfactory.

- Domain Specific Indexing and Retrieval. High-level indexing and retrieval is in great need in many applications. However, as mentioned above, it is technically challenging or involves extensive manual efforts. Alternatively, domain specific indexing and retrieval can use the high level structure of video or *a priori* knowledge to assist the extraction of video features and/or semantic meanings. These techniques are effective in their intended domain of application.

- Object-level Indexing and Retrieval. In this method, the salient video objects are used to represent the spatio-temporal characteristics of video clips [6][12]. The motivation of this approach is that any given scene is generally a complex collection of objects. Thus the location and physical qualities of each object, as well as their interaction with each other, define the content of the scene and the extracted object can serve as a valuable visual index cue.

- Low-level Indexing and Retrieval. Such techniques provide access to video based on properties like color and texture. The driving force behind this group of techniques is to organize the features based on some distance metric and to use similarity based matching to retrieve the video. Their primary limitation is the lack of semantics attached to the features.

As discussed in [145], from the user's point of view, there are mainly two kinds of video retrieval demands: visual query and concept query. Visual query refers to the cases when users want to find video shots that are visually similar to a given example, which can be realized by directly comparing low level visual features of video shots or their key frames. Obviously, this type of query request can be well supported by a low-level indexing scheme. However, generally users are more interested in concept query, that is, to find video shots by the presence of specific objects or events. Although a number of researches have been conducted to model and retrieve the video data based on objects [12][127], the performance is still largely limited by the difficulties of automatic object extraction, tracking and recognition. There are also considerable approaches which attempt to capture video events, especially for specific application domains such as sports videos [67], traffic videos [65], etc. Specifically, in most existing works, event detection is normally carried out in a two-step procedure [67]. In the first step, low-level descriptors are extracted from the video documents to represent the low-level information in a compact way. Then in the second step, a decision-making algorithm is used to explore the semantic index from the low-level descriptors. For instance, in the domain of sports videos, depending on the types of low-level features extracted and utilized for event detection, the frameworks can be classified into two categories: unimodal (using only the visual [32], auditory [93], or textual modality [140]) and multimodal [2] [126]. The multimodal approach attracts growing attention nowadays as it captures the video content in a more comprehensive manner. In terms of the decision-making algorithms, the Markov model-based techniques have been extensively studied, including the Hidden Markov Model (HMM) [132], Controlled Markov Chain (CMC) [67], etc. to model the temporal relations among the frames or shots for a certain event. Another type of heuristic method uses a set of heuristic rules that are derived from the domain knowledge to map the feature descriptors to events [120][147]. In addition, a multimedia data mining approach was presented in our earlier work [11][14] to mine the high-level semantics and patterns from a large amount of multimedia data. However, the semantic gap issue remains a major obstacle and most of approaches require vast amounts of manual efforts or domain knowledge.

CHAPTER 3

Overview of the Framework

The advances in data capturing, storage, and communication technologies have made vast amounts of multimedia data available to consumer and enterprise applications. However, the tools and techniques are still limited in terms of describing, organizing, and managing multimedia data. In this dissertation, an integrated multimedia data management and retrieval framework will be proposed with the focus to address the semantic gap and perception subjectivity issues and to facilitate the effective data organization. Fig. 3.2 shows the proposed framework. As can be seen, it consists of three major components: data representation, indexing and retrieval. These three components are integrated closely and act as a coherent entity to support the essential functionalities of an MMDBMS. Specifically, data representation serves as the basis for effective indexing and retrieval. To bridge the semantic gap, the data representation generally consists not only of the low-level media features, but also mid-level and knowledge-assisted descriptors. Indexing is then performed upon the data representation to ensure a fast searching and retrieval mechanism of the media objects.

As discussed in Section 1.2, image database and video database serve as the test beds for this framework. Therefore, the framework will be detailed for them separately for the purpose of clarity. However, as a video stream consists of a consecutive sequence of image and sound data with temporal constraints, many techniques adopted for image indexing and retrieval can serve as the basis for video database.

3.1 Image Database

3.1.1 Image Data Representation

Global features (i.e., image-level features), such as color and texture, have long been used for image content analysis and representation. Recently, with the assumption that human discernment of certain visual contents could be potentially associated with the semantically meaningful object(s) in the image, region-level or object-level features should also be explored, where an image segmentation process must be performed beforehand to decompose the images into a set of homogeneous regions. In this proposed framework, an unsupervised image segmentation method called WavSeg [141] is adopted to partition the image. Then both image-level and object-level color and texture features are extracted.

Though the extraction of the above-mentioned features is the basis for image content analysis, these features alone are inadequate to represent the image semantic meanings. To tackle this issue, in this proposal, a semantic network framework is proposed, which is constructed from the viewpoint of long-term learning based on the concept of Markov Model Mediator (MMM) [99]. Different from the general Relevance Feedback approach where the user's feedback is solely used to customize the current query and then discarded, the basic idea of semantic network is that such feedback contains valuable semantic information and should be accumulated for a stochastic modeling scheme to capture the semantic concepts from the viewpoint of the majority users. Therefore, to construct the semantic network, the relevance feedback information is accumulated in the database log and acts as the training data set. A training process is thus applied upon the training data to stochastically model the semantic relationships among the images. As mentioned earlier, such a network is useful because image retrieval is basically a process to explore the relationships between the query image and the other images in the database. Since the network is built by using the accumulated relevance feedback information provided by a group of users, it possesses the capability of modeling the users' general concepts in regard to the images' semantic meanings.

3.1.2 Image Data Organization

Efficiency is another important issue in CBIR. For a large image database, the traditional retrieval methods such as sequential searching do not work well since they are time expensive. Two approaches have been widely used for the sake of retrieval efficiency: 1)

Feature space reduction and search space reduction; 2) Indexing and data structures for organizing image feature vectors. For feature reduction, principal component analysis (PCA) and wavelet transform are two commonly used techniques to generate compact representations for original feature space. For search space reduction, there are various pre-filtering processes [75], such as filtering with structured attributes, methods based on triangle inequality, and filtering with color histograms [42].

In terms of developing the appropriate indexing techniques and data structures to speed up the image search process, it remains an open issue. Many data structures, approaches and techniques have been proposed to manage an image database and hasten the retrieval process, such as VA-file [128], M-tree [24], and MB+-trees [29]. The QBIC system [35], for instance, uses the pre-filtering technique and the efficient indexing structure like R-trees to accelerate its searching performance. As another example, the ImageScape system [68] uses k-d tree as its indexing structure. A survey of the techniques and data structures for efficient multimedia retrieval based on similarity was given in [75]. Most of the existing works on data indexing and data structures are conducted at a single data set level based on low-level features. However, to address the semantic gap issue, there's a strong need to index the image not only based on its low-level features (low-level indexing) or object-level features (object-level indexing), but also on the high-level representations. In contrast, clustering [43] is one of the most useful knowledge discovery techniques for identifying the correlations in large data sets. There are different types of clustering algorithms in the literature such as the partitioning clustering algorithms [79], hierarchical clustering methods where a representative of the data set is computed based on their hierarchical structures [110], and syntactic methods which are based on the static structure of the information source [61]. One of the disadvantages of the syntactic method is that it ignores the actual access patterns. Another type of method collects the statistics pertaining to the access patterns from feedback logs and conducts partitioning based on the statistics [101], which will be extended in this dissertation to organize the image data in the database. In addition, most of the existing work concentrates on data indexing and data clustering schema at a single database level, which is not sufficient to meet the increasing demand of handling efficient image database retrieval in a distributed environment, in which the query process may be carried out across several image databases residing in distributed places and the query results may come from different databases. A database clustering approach will be of great help in this manner and will also be studied.

3.1.3 Image Retrieval

To effectively tackle the perception subjectivity issue, a probabilistic semantic network-based image retrieval framework incorporated with Relevance Feedback (RF) technique is proposed [100], which not only takes into consideration the low-level image content features, but also fully utilizes the relative affinity measurements captured in the semantic network. Therefore, instead of starting each query with low-level features as conducted by most of the existing systems, the users' general perceptions are gradually embedded into the framework to improve the initial query results. To serve for individual user's query interests, the semantic network is intelligently traversed based on user's feedback and the query results are refined accordingly. This framework thus possesses the capability of capturing the general user concepts and meanwhile adjusting to the perception with regard to a specific user. One potential limitation is that this framework models user's perception at the image-level and has difficulties in propagating the feedback information across the query sessions toward the region or object level.

In order to dynamically discover the object in the image that is the focus of a user's attention, an advanced CBIR system called MMIR is proposed by further extending the above-mentioned framework [7]. In brief, the MMIR system utilizes the MMM mechanism to direct the focus on the image level analysis together with the Multiple Instance

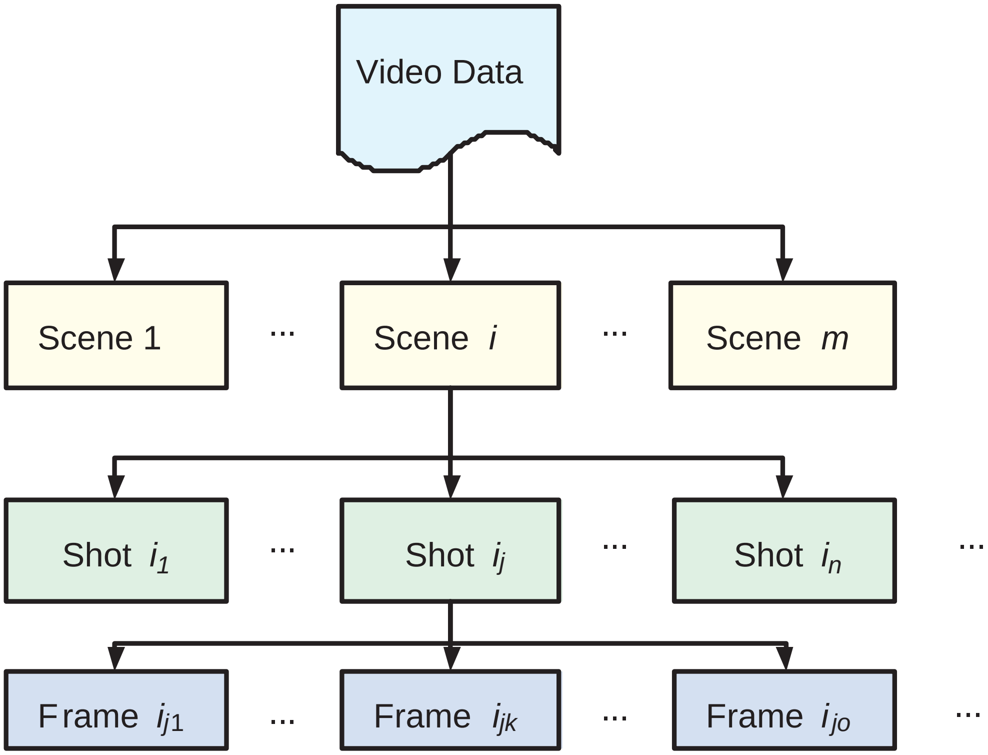


Figure 3.1: General video structure.

Learning (MIL) technique (with the Neural Network technique as its core) for real-time capturing and learning of the object-level semantic concepts with the facilitation of user feedback. In addition, from a long-term learning perspective, the user feedback logs explored by the semantic network are used to speed up the learning process and to increase the retrieval accuracy for a query.

3.2 Video Database

3.2.1 Video Data Representation

Generally speaking, video data can be modeled hierarchically as shown in Fig. 3.1. Here, a video clip is composed of a set of video scenes, which can be defined as a collection of semantically related and temporally adjacent shots. A shot in turn consists of an unbroken sequence of frames taken from one camera. Since shot is widely considered as a self-contained unit, in this dissertation, the shot-based approach is adopted in terms of modeling and mining videos in a video database system. Therefore, a shot-boundary detection approach is first applied to segment the videos into a set of meaningful and manageable units. Then the visual/audio features are extracted for each shot at different granularities. Specifically, shot-level features are obtained by averaging the feature values within the shot range. In addition, the shot is also abstracted and represented by key frame(s) whose content is also analyzed and features are extracted correspondingly. Note that as illustrated in Fig. 3.2, the feature extraction and object segmentation techniques developed for image content analysis are readily used in the video shot detection process and visual feature extraction from shots and key frames. Meanwhile, audio features in time domain and frequency domain are explored.

Although the low-level features are captured in multiple channels (or called multi-modal features) and contain more complete video information in comparison to the uni-modal approach, due to the complexity of video contents, they alone are generally not sufficient to deliver comprehensive content meanings. Therefore, two main approaches are proposed in this dissertation, namely mid-level representation extraction and automatic knowledge discovery (to construct knowledge-assisted representation), to enrich the video data representation and to bridge the semantic gap.

Mid-level representations are deduced from low-level features and are motivated by high-level inference or *a priori* knowledge. Taking the soccer videos as the test bed, four mid-level features describing camera view types, field ratio, level of excitement, and corner view types are proposed. Among them, the first three features are generic descriptors for field-sports videos as they are not game specific. In contrast, the corner view type descriptor is semi-generic because while it is useful in identifying corner events (corner kicks, line throws from the corner, free kicks close to the penalty box, etc.) in soccer videos, it is a less important indicator of events in other types of field-sports.

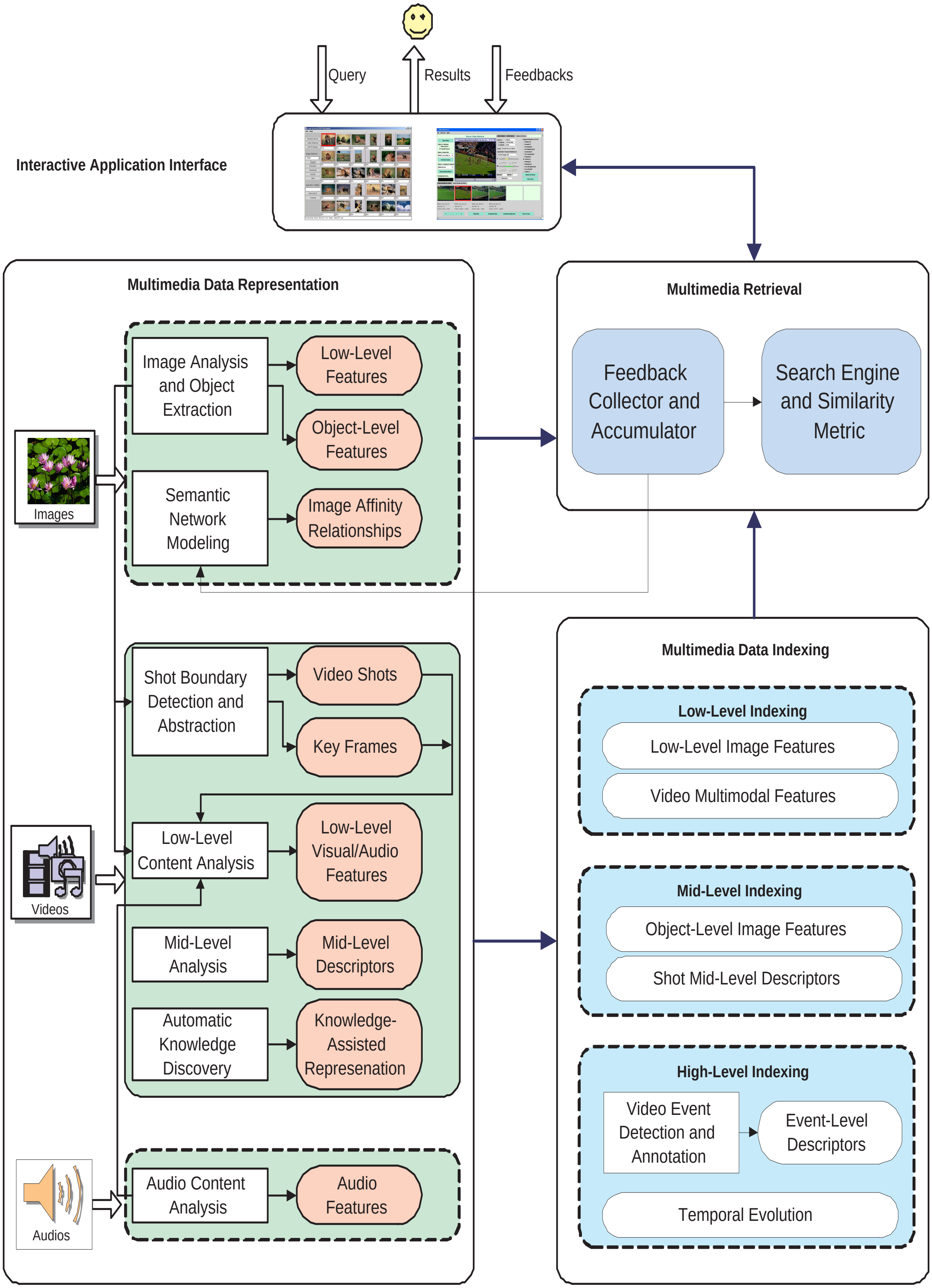

Query
Results
Feedbacks
Interactive Application Interface
Multimedia Data Representation
Image Analysis and Object Extraction
Low-Level Features
Object-Level Features
Semantic Network Modeling
Image Affinity Relationships
Images
Shot Boundary Detection and Abstraction
Video Shots
Key Frames
Low-Level Content Analysis
Low-Level Visual/Audio Features
Videos
Mid-Level Analysis
Mid-Level Descriptors
Automatic Knowledge Discovery
Knowledge-Assisted Representation
Audio Content Analysis
Audio Features
Audios
Multimedia Retrieval
Feedback Collector and Accumulator
Search Engine and Similarity Metric
Multimedia Data Indexing
Low-Level Indexing
Low-Level Image Features
Video Multimodal Features
Mid-Level Indexing
Object-Level Image Features
Shot Mid-Level Descriptors
High-Level Indexing
Video Event Detection and Annotation
Event-Level Descriptors
Temporal Evolution

Figure 3.2: Overview of the proposed framework.

As an alternative approach, automatic knowledge discovery algorithms are proposed to intelligently model knowledge-assisted data representation and to relax the framework's dependence on the domain knowledge and human efforts by fully exploring the temporal evolution and context information in the video data. Two advanced techniques are developed for this purpose, namely temporal segment analysis [8] and temporal association mining.

In temporal segment analysis, a novel time window algorithm is conducted to automatically search for the optimal temporal segment and its associated features that are significant for characterizing the events. Then a temporal pattern clustering process is performed for data reduction to boost the event detection performance. One limitation of this approach is that although a significant time window for certain events can be effectively identified, it has difficulties to systematically capture and model the characteristic context from the time window. The reason lies in the fact that such important context might occur at uneven inter-arrival times and display at different sequential orders. Therefore, the concept of temporal segment analysis is further extended and a temporal association mining scheme is applied to not only explore the characteristic temporal patterns but also offer an intelligent representation of such patterns. The basic idea is that the problem of finding temporal patterns can be converted so as to find adjacent attributes which have strong associations with (and thus characterize) the target event. Therefore, association rule mining provides a possible solution since the inference made by association rule suggests a strong co-occurrence relationship between items [115]. However, the problem of temporal pattern discovery for video streams has its own unique characteristics, which differs greatly from the traditional association rule mining and is thus tackled by the proposed hierarchical temporal association mining framework.

3.2.2 Video Indexing and Retrieval

If we perceive a video clip as a document, video indexing can then be analogous to text document indexing where a structural analysis is first performed to decompose the document into paragraphs, sentences, and words, before building indices. In general, a video can be decomposed into scenes, shots, frames, etc. In this work, to facilitate fast and accurate content access to video data, the video document is segmented into shots as mentioned earlier. Its shot-level features and key frame features can be acted as its index entry. Such indexing is mainly used to support the visual queries, such as key frame query. That is, the system will retrieve the video shots visually similar to a given example by extracting its low-level visual features and directly comparing them with the features of the key frames stored in the database. In this case, the query mechanism and search engine applied in the image database can be readily used, with the only exception that instead of returning a static image, a corresponding shot is displayed.

In addition, the users are generally more interested in concept query, that is, to find video shots by the presence of specific events, as discussed in Section 2.2.3. In response to such requests, important activities and events are detected in this work by applying the data classification algorithm on a combination of multimodal mid-level descriptors (or knowledge-assisted data representation) and low-level features. Specifically, the decision tree learning algorithm is adopted for this purpose as it is mathematically less complex and possesses the capability of mapping data representation to high-level concepts [97]. This framework has been tested using soccer videos with different production styles and from different video sources. In addition, this framework is also extended to detect important concepts (i.e., high-level semantic features), such as "commercial," "sports," from TRECVID news videos [122]. It is worth noting that TRECVID program was led by the National Institute of Standards and Technology to provide a benchmark for multimedia research by offering common data set and common evaluation procedure.

Such event/concept labels can thus be tagged to the shots for high-level video indexing or video annotation to support concept query.

CHAPTER 4

Data Management and Retrieval for Image Database

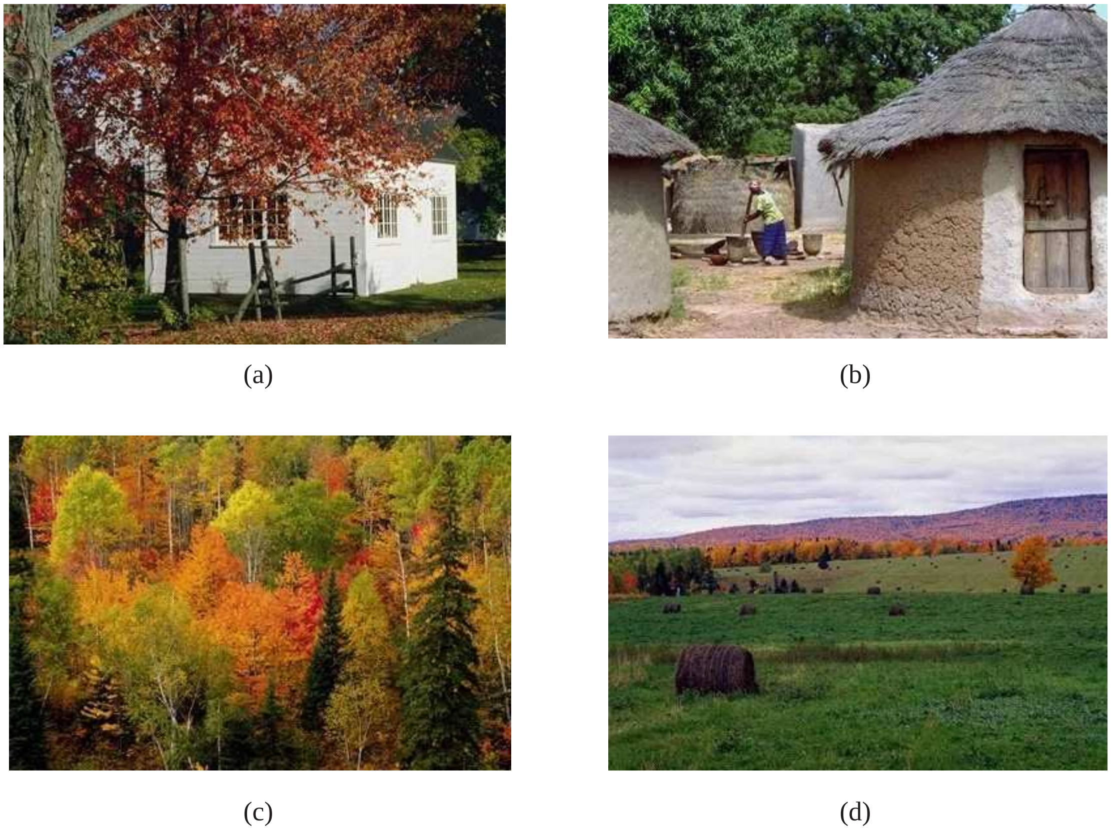
The explosive growth of image data has made efficient image indexing and retrieval mechanism indispensable. As mentioned in Section 1.1, semantic gap and perception subjectivity issues are two of the major bottlenecks for CBIR systems. In this chapter, these issues are addressed from the perspectives of both image data representation and image retrieval processes.

4.1 Image Data Representation

It is widely accepted that the major bottleneck of CBIR systems is the large semantic gap between the low-level image features and high-level semantic concepts, which prevents the systems from being applied to real applications [50]. Therefore, in terms of image data representation, besides the structured description of visual contents, a semantic network is constructed to capture the relative affinity measurement among the images, which can serve as an essential data representation to bridge the semantic gap.

The perception subjectivity problem poses additional challenges for CBIR systems. In other words, as illustrated in Fig. 1.2 (for better understanding, this figure is inserted again in this Section) and discussed in Section 1.1, in viewing the same image (e.g., Fig. 4.1(a)), different users might possess various interests in either a certain object (e.g., the house, the tree, etc.) or the entire image (e.g., a landscape during the autumn season). In addition, even the same user can have different perceptions towards the same image at various situations and with different purposes. Therefore, features from both the image and object levels are required to support the multi-level query interests.

The main focus of this study is to propose effective approaches to mitigate the above-mentioned issues instead of exploring the most appropriate features for image indexing and retrieval. Therefore, in the next two subsections, a brief introduction will be given for feature extraction followed by a detailed discussion about the semantic network.



(a)

(b)

(c)

(d)

Figure 4.1: Example images

4.1.1 Global Features

Two groups of global features, color and texture, are extracted at the image level.

- Color Feature. Since the color feature is closely associated with image scenes and it is more robust to changes due to scaling, orientation, perspective and occlusion of images, it is widely adopted in a CBIR system for its simplification and effectiveness. The HSV color space is used to obtain the color feature for each image in this study for the following two reasons: 1) HSV color space and its variants are proven to be particularly amenable to color image analysis [23], and 2) it was shown in the benchmark results that the color histogram in the HSV color space has the best performance [77]. Also as discussed in [37], though the wavelength of visible light ranges from 400 to 700 nanometers, the colors that can be named by all the cultures are generally limited to be around 11. Therefore, the color space is quantized using color categorization based on H, S, V value ranges and 13 representative colors are identified [7]. Besides black and white, ten discernible colors ('red,' 'red-yellow,' 'yellow,' 'yellow-green,' 'green,' 'green-blue,' 'blue,' 'blue-purple,' 'purple,' and 'purple-red') are extracted by dividing the Hue into five main color slices and five transition color slices. Here, each transition color slice like 'red-yellow,' 'yellow-green,' is considered between two adjacent main color slices. In addition, a new category 'gray' is added for the remaining value ranges. Colors with the number of pixels less than 5% of the total number of pixels are regarded as non-important and the corresponding positions in the feature vector have the value 0. Otherwise, the corresponding percentage of that color component will be used.

- Texture Feature. Texture is an important cue for image analysis. It has been shown in a variety of studies [107][118] that characterizing texture features in terms of structure, orientation, and scale fits perfectly with the models of human perception. A number of texture analysis approaches have been proposed. In this study, a one-level wavelet transformation using Daubechies wavelets is used to generate the horizontal detail sub-image, the vertical detail sub-image, and the diagonal detail sub-image. The reason for selecting Daubechies wavelet transform lies in the fact that it was proven to be suitable for image analysis. For the wavelet coefficients in each of the above three subbands, the mean and variance values are collected, respectively. Therefore, six texture features are extracted.

In summary, the extraction process for global feature extraction is relatively straightforward where an image is considered as a whole and a vector of 19 features (13 color features and 6 texture features) is generated in this study as discussed above. Note that for simplicity, it is assumed that the color and texture information are of equal importance such that the values of the color features should be equal to those of the texture features. Therefore, a feature normalization process is conducted. As a result, the sum of all feature values for a given image is 1, with both the sum of all the color features and that of texture features being 0.5.

4.1.2 Object-level Features

As far as the region level features are considered, an image segmentation process needs to be carried out beforehand.

In this study, the WavSeg algorithm proposed in our earlier work [141] is applied to partition the images. In brief, WavSeg adopts a wavelet analysis in concert with the SPCPE algorithm [19] to segment an image into a set of regions. The problem of segmentation is converted to the problem of simultaneously estimating the class partition and the parameter for each class and is addressed in an iterative process. By using Daubechies wavelets, the high-frequency components will disappear in larger scale subbands and the possible regions will be clearly evident. Then by grouping the salient points from each channel, an initial coarse partition is obtained and passed as the input to the SPCPE segmentation algorithm, which has been proven to outperform the random initial partition-based SPCPE algorithm. In addition, this wavelet transform process can actually produce region-level texture features together with the extraction of the region-of-interest within one entry scanning through the image data. Once the region information becomes available, the region-level color features can be easily extracted.

4.1.3 Semantic Network

To tackle the semantic gap issue, a probabilistic semantic network is proposed based on the concept of Markov Model Mediator (MMM), a probabilistic reasoning model that adopts the Markov model framework and the concept of mediators. The Markov model is one of the most powerful tools available for scientists and engineers to analyze complicated systems, whereas a mediator is defined as a program to collect and combine information

Table 4.1: The relative affinity matrix $\mathcal{A}$ of the example semantic network.

	Img 1	**Img 2**	**Img 3**	**Img 4**	**...**	**Img m**	**...**	**Img N**
Img 1	$a_{1,1}$	$a_{1,2}$	0	0	...	$a_{1,m}$	...	0
Img 2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	...	0	...	0
Img 3	0	$a_{3,2}$	$a_{3,3}$	0	...	$a_{3,m}$	...	0
Img 4	0	$a_{4,2}$	0	$a_{4,4}$	...	0	...	0
...	...	...	...	...	...	...	...	...
Img m	$a_{m,1}$	0	$a_{m,3}$	0	...	$a_{m,m}$	...	0
...	...	...	...	...	...	...	...	...
Img N	0	0	0	0	...	0	...	$a_{N,N}$

from one or more sources, and finally yield the resulting information [130]. Markov models have been used in many applications. Some well-known examples are Markov Random Field Models in [36], and Hidden Markov Models (HMMs) [92]. Some research work has been done to integrate the Markov model into the content-based image retrieval. Lin et al. [71] used a Markov model to combine spatial and color information. In their approach, each image in the database is represented by a pseudo two-dimensional hidden Markov model (HMM) in order to adequately capture both the spatial and chromatic information about that image. In [83], a hidden Markov model was employed to model the time series of the feature vector for the cases of events and objects in their probabilistic framework for semantic level indexing and retrieval. In brief, the MMM mechanism contains two major parameters, namely the relative affinity matrix $\mathcal{A}$ and the feature matrix $\mathcal{B}$, to represent both the semantic network and the low-level features for the images in the database. Note that the MMM mechanism directs the focus on the image level analysis; therefore matrix $\mathcal{B}$ contains global features as introduced in Section 4.1.1 and matrix $\mathcal{A}$ (the semantic network) is constructed [100] as follows.

Assume N is the total number of images in the image database and $I = \{i_1, i_2, \dots, i_N\}$ is the image set. The semantic network is modeled by the relative affinity matrix $\mathcal{A}$, where $\mathcal{A} = \{a_{m,n}\}$ $(1 \leq m, n \leq N)$ denotes the probabilities of the semantic relationships among the images based on users' preferences, and the relationships of the images in the

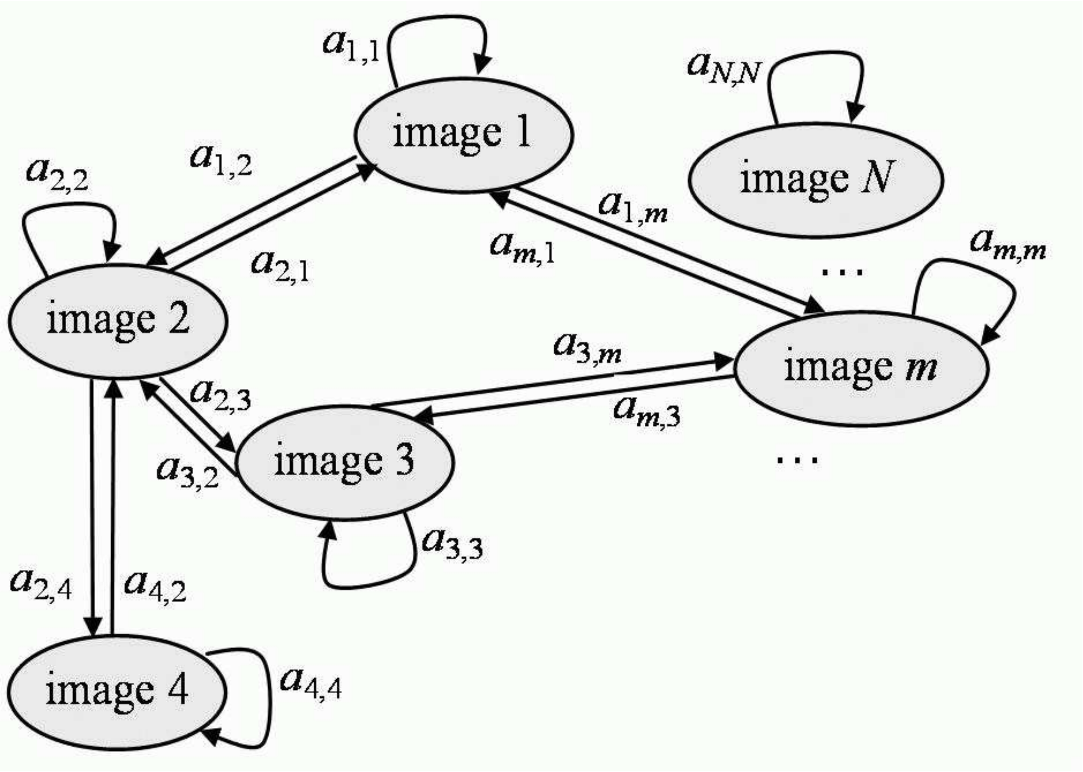


Figure 4.2: Probabilistic semantic network.

semantic network are represented by the sequences of the states (images) connected by transitions. Fig. 4.2 shows an example of the semantic network, where the lines with zero probabilities are omitted. Table 4.1 shows the corresponding $\mathcal{A}$ matrix.

In the network, two different kinds of relationships are defined between two images:

1. Directly related (R_D)

   $i_m \; R_D \; i_n \Leftrightarrow a_{m,n} \neq 0$ where $i_m, i_n \in I$, $a_{m,n} \in \mathcal{A}$

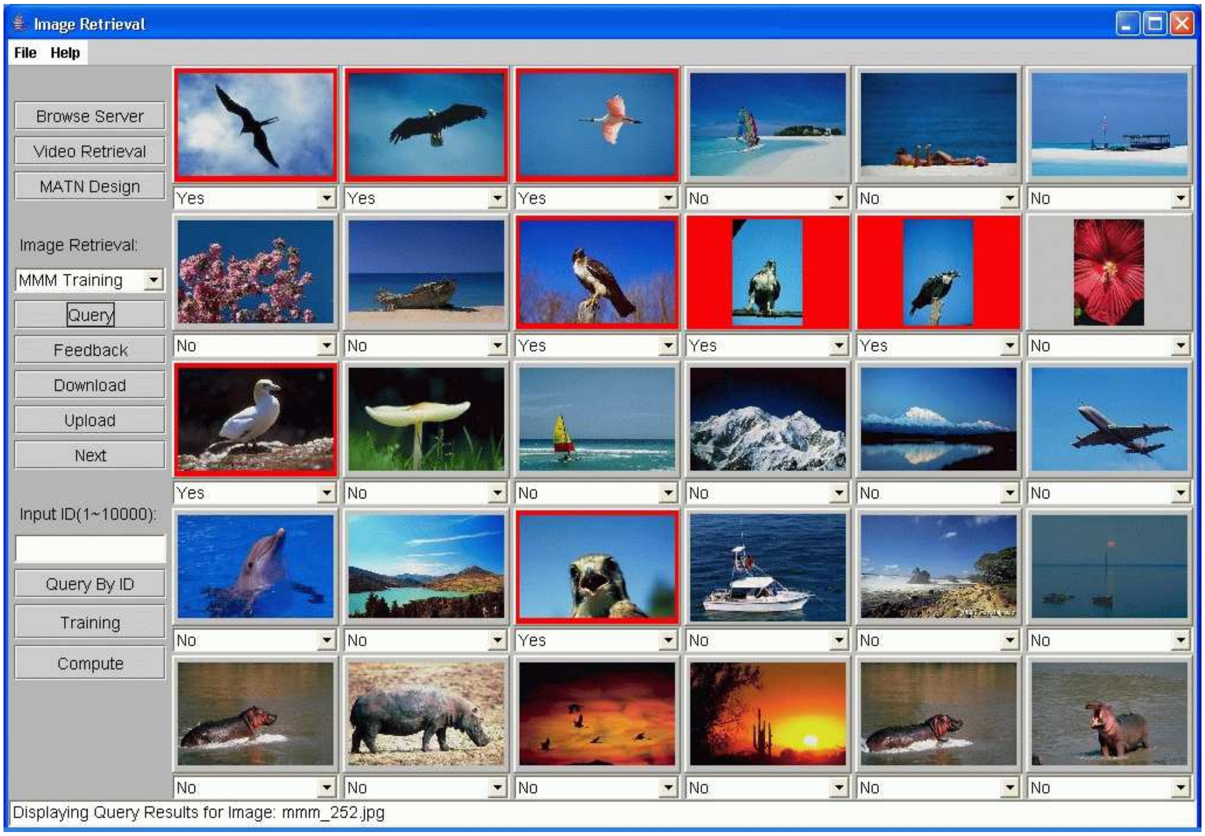
   For example: $i_1 \; R_D \; i_2$, $i_2 \; R_D \; i_3$, etc.

2. Indirectly related (R_I)

   $i_m \; R_I \; i_n \Leftrightarrow ((a_{m,n} = 0) \wedge (\exists i_x \in I \Rightarrow a_{m,x} \neq 0 \wedge a_{x,n} \neq 0))$ where $i_m, i_n, i_x \in I$, $a_{m,n}, a_{m,x}, a_{x,n} \in \mathcal{A}$, and $m \neq n$

   For example: $i_1 \; R_I \; i_3$, $i_1 \; R_I \; i_4$, etc.

In other words, R_D is the relationship between two directly linked images, while R_I exists between two images that are connected to a common image. For the purpose of constructing the semantic network, a set of training data is needed.



Displaying Query Results for Image: mmm_252.jpg

Figure 4.3: The interface of the training system.

Training Data Set

To construct the semantic network, a training data set is required to generate the probabilistic semantic relationships among the images. The source of the training data set is the log of user access patterns and access frequencies on the image data. Access patterns denote the co-occurrence relationships among the images accessed by the user queries, while access frequencies denote how often each query was issued by the users.

To collect the user access patterns and access frequencies, an image retrieval system implemented earlier by our research group [18] is adopted in this framework. Fig. 4.3

shows the system interface. In brief, the training process is described as follows: The user first selects one query image. After the “Query” button is clicked, a query message is sent to the server through UDP. The query results are sent back after the server fulfills the query process. It is worth mentioning that for training purpose, any available image retrieval methods can be implemented on the server side. Upon receiving the results, the user selects the images that he/she thinks are related to the query image by right-clicking on the image canvases, and clicks the “Feedback” button to send the feedback back to the server. When the server receives and identifies this feedback message, it updates the user access patterns and access frequencies accordingly. Then the user can continue the training process or exit. Detailed information can be found in [18]. In this study, a group of users were asked to randomly issue queries and select positive and negative examples from the results for each query. The positive examples selected in each query are said to have the co-occurrence relationships with each other. Intuitively, they are semantically related. In addition, the more frequently two images are accessed together, the more closely they are related. It is worth mentioning that such a process is actually supported by most of image database systems which offer interactive user interfaces for users to provide feedback on the query results (and possibly refine the results accordingly using Relevance Feedback principle). In real applications, by accumulating feedback in the log mechanism, the semantic network concept can be readily applied to these systems for performance improvement.

The formal definition regarding the training data set is given as follows:

Definition 4.1. Assume N is the total number of images in the image database and a set of queries $Q = \{q_1, q_2, ..., q_{nq}\}$ were issued to the database in a period of time. The training data set consists of the following information:

- Let $use_{m,k}$ denote the usage pattern of image m with respect to query q_k per time period, where the value of $use_{m,k}$ is 1 when m is accessed by q_k and zero otherwise.

Table 4.2: The query access frequencies ($access_k$) and access patterns ($use_{k,m}$) of the sample images.

Query	$access_k$	**Img (a)**	**Img (b)**	**Img (c)**	**Img (d)**	**...**
q_1	4	1	1	0	...	...
q_2	1	0	1	1	...	...
q_3	$access_3$	0	0	0	...	...
...	...	...	...	...	...	...

- The value of $access_k$ denotes the access frequency of query q_k per time period.

Table 4.2 gives some example queries issued to the image database with their corresponding access frequencies. The access patterns of the four sample images in Fig. 4.1 versus the example queries are also shown in Table 4.2. In this table, the entry (k, m) = 1 indicates that the m^{th} image is accessed by query q_k. For example, suppose q_1 is a user-issued query related to retrieving images containing country house scenes. Img (a) and (b) are accessed together in q_1, with their corresponding entries in the access pattern matrix having the value 1. Let q_2 denote a query related to the concept of '*trees*', then Img (a) and Img (c) will probably be accessed together in this query. However, since most users regard Img (a) as a country house scene more than a '*trees*' scene, the access frequency of q_1 is larger than that of q_2. Consequently, after the system training, Img (b) is more likely to be retrieved than Img (c), given that Img (a) is selected as the query image. Thus, the users' subjective concepts about the images are captured by the pair of user access patterns and user access frequencies.

Construct the Semantic Network

Based on the information in the training data set, the semantic relationships can be captured among the images in the database and construct the semantic network. In order to capture the semantic relationships among all the images, an assisting matrix AFF is defined, which is constructed by having the $aff_{m,n}$ to be the relative affinity relationship between two images m and n using the following definition.

Definition 4.2. The relative affinity measurement ($aff_{m,n}$) between two images m and n ($1 \leq m, n \leq N$) indicates how frequently these two images are accessed together, where

$$aff_{m,n} = \sum_{k=1}^{nq} use_{m,k} \times use_{n,k} \times access_k \tag{4.1}$$

Let N be the total number of the images in the database. The matrix $\mathcal{A}$ is initialized by having $a_{m,n}$ be the element in the $(m,n)^{th}$ entry in $\mathcal{A}$, where

$$a_{m,n} = 1/N \tag{4.2}$$

Then $\mathcal{A}$ is constructed via the following equation.

$$a_{m,n} = \begin{cases} \frac{aff_{m,n}}{\sum_{k \in d} aff_{m,k}} & \text{if } \sum_{k \in d} aff_{m,k} \neq 0 \\ a_{m,n} & \text{otherwise.} \end{cases} \tag{4.3}$$

From the above equations, it can be seen that each row m in the $\mathcal{A}$ matrix represents the R_D relationship between the image i_m and the other images, while the whole $\mathcal{A}$ matrix contains information of both kinds of relationships among the images in the database. The values in $\mathcal{A}$ are then used to construct the semantic network.

For the sake of efficiency, during a training period, the training system only collects all the user access patterns. Once the number of records reaches a threshold (e.g., 500), an update of $\mathcal{A}$ matrix is triggered automatically. All the computations are done off-line.

Moreover, instead of using the whole $\mathcal{A}$ matrix, in the retrieval process only the R_D relationship between the query image and the other images together with the low-level features are used to generate the initial query results. In other words, for a specific query image i_m ($i_m \in I$), only the m^{th} row in matrix $\mathcal{A}$ (denoted as A_m) is applied in order to reduce the computational load and I/O cost.

4.2 Content-based Image Retrieval

In our earlier studies, the MMM mechanism was applied to multimedia database management [104][105] and document management on the World Wide Web (WWW)

[101][102]. MMM is used for content-based image retrieval and the preliminary results were presented in [99], where the MMM mechanism functions as both the searching engine and the image similarity arbitrator for image retrieval. However, MMM considers only the direct relationship between the query image q and the other images in the database and does not support real-time query refinement based on user feedback. Therefore, in the following two sections, two extended frameworks are presented to support real-time query refinement and to explore user perceptions at both the image-level and object-level.

4.2.1 Probabilistic Semantic Network-Based Image Retrieval

As discussed in Section 2.1.3, RF aims to adjust to a specific user's query perception according to the user's feedback. However, the RF technique is often inadequate in learning the concepts [55] due to the limited power of the low-level features in representing high-level semantics. To address this issue, a framework that performs relevance feedback on both the images' low-level features and the semantic contents represented by keywords was proposed [76]. In their work, a semantic network was constructed as a set of keywords linked to the images in the database. Though the retrieval accuracy is improved by using this approach, extra effort is required to label the images manually with the keywords. To overcome such limitations, in this section, a probabilistic semantic network-based image retrieval framework, called MMM_RF, is proposed [100], which employs both relevance feedback and Markov Model Mediator (MMM) mechanism for image retrieval. The low-level global features and semantic network representation extracted in Section 4.1.1 and Section 4.1.3 respectively, are fully utilized. This proposed framework can support both accumulative learning to capture general user concepts and on-line instance learning for individual user interests.

The architecture of the proposed framework is shown in Fig. 4.4. As can be seen from this figure, a training process is used to construct the semantic network off-line as discussed in Section 4.1.3. Then image retrieval is conducted by utilizing both the

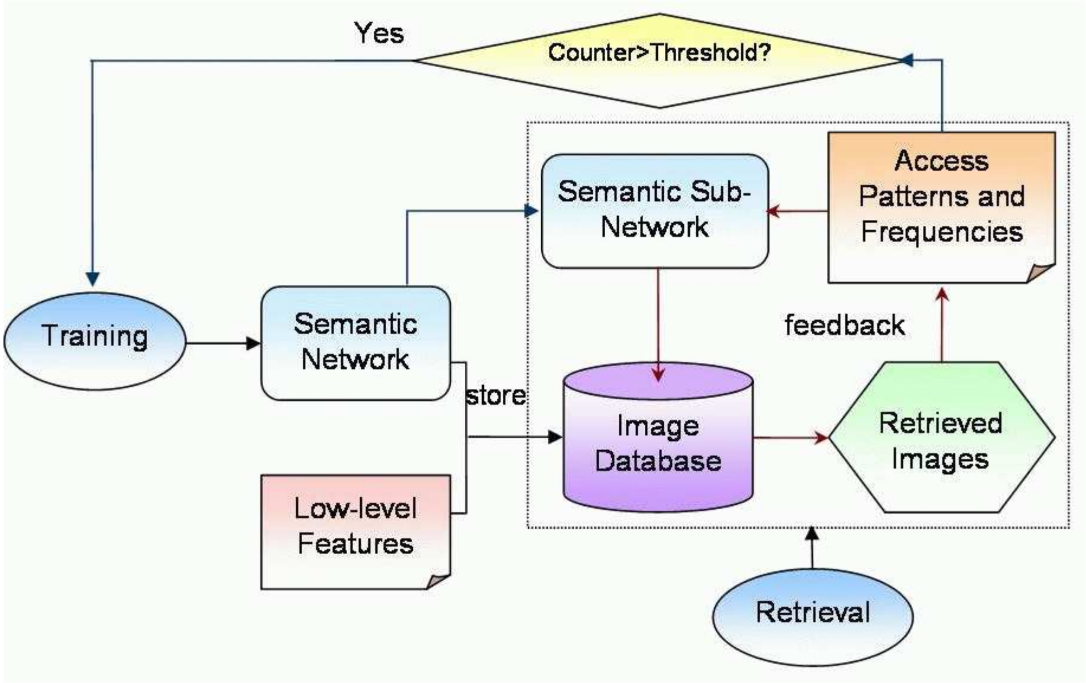


Figure 4.4: Framework architecture.

low-level features and the semantic network captured in the feature matrix $\mathcal{B}$ and the relative affinity matrix $\mathcal{A}$, respectively. A feedback process refines the current retrieval results by updating the temporary semantic subnetwork. Meanwhile, the user feedback is collected continuously as the training data for subsequent updates of the whole semantic network. A discussion about on-line retrieval is detailed in the following subsection.

Refine the Semantic Subnetwork On-Line

As mentioned above, the relative affinity matrix $\mathcal{A}$ is obtained based on the feedback provided by various users on different kinds of queries. Therefore, matrix $\mathcal{A}$ represents the general user concepts and can help to achieve better query results. However, in the retrieval process, different users may have different concepts about the query images. Therefore, in addition to supporting accumulative learning, the system also needs to support instant learning which enables the query refinement for individual users on the fly. Moreover, it is most likely that the query images chosen by the users have no existing R_D or R_I relationships in the current semantic network. In this subsection, a refinement method for the semantic subnetwork is proposed to solve these problems based on user feedback. Note that as discussed earlier, such refinement is conducted on the temporary subnetwork only for the sake of system efficiency and avoiding the bias caused by a single user.

For a query image i_m ($i_m \in I$), the user can choose to accept the initial query results obtained by using the general user concepts, or to provide feedback via indicating the positive and negative examples. The access patterns can be obtained based on these positive and negative examples, as mentioned in the previous subsection. Such access patterns are then used to update the $\mathcal{A}$ matrix to further improve the initial query results. More importantly, the user specified R_D relationship $\mathcal{A}^{'}_{m} = [a^{'}_{k}]$ $(1 \leq k \leq N)$ between i_m and other images can be obtained by using Eqs. (4.1) and (4.3) with parameter m fixed and n varied from 1 to N. Let vector $V_m = [v_j]$ $(1 \leq j \leq N)$ denote the information of both R_D and R_I relationships between i_m and other images. Table 4.3 shows the steps to calculate it. The idea is quite straightforward. As we know, each $a^{'}_{n_i}$ in $\mathcal{A}^{'}_{m}$ represents the R_D probability from i_m to i_{n_i}, while each $a_{n_i,j}$ in $\mathcal{A}_{n_i}$ denotes the R_D probability from i_{n_i} to i_j. Therefore, the R_D probability from i_m to i_j connected by a common image i_{n_i} can be obtained by $a^{'}_{n_i} \times a_{n_i,j}$. A pair of images can be indirectly related to each other via multiple paths in the semantic network. For example, Img 1 and Img 3 are indirectly related via two different ways – one through Img 2 and one through Img m (as shown in Fig. 4.2). In such a situation, the maximal probability is kept since this maximal probability indicates the actual degree of their semantic relationship. It is worth mentioning that normally the set of non-zero items in $\mathcal{A}^{'}_{m}$ is quite small, so the algorithm is efficient in terms of space and time.

Stochastic Process for Information Retrieval

In this subsection, a stochastic retrieval process is defined to calculate the edge weights among the images utilizing both the low-level features and the semantic network. Assume

Table 4.3: Capture R_D and R_I relationships between i_m and other images.

```
1. Obtain non-zero items $[a'_{n_1}, a'_{n_2}, ..., a'_{n_T}]$ in $A'_m$, where $T$ is the total number of
   images which have non-zero $R_D$ relationships with $i_m$ $(1 \leq T \leq N)$
2. For each $a'_{n_i}$ $(1 \leq i \leq T)$, get its corresponding $A_{n_i} = [a_{n_i,j}]$ $(1 \leq j \leq N)$ from
   matrix $\mathcal{A}$
3. Normalize each $A_{n_i}$ as:
   if $(\max(A_{n_i}) \neq \min(A_{n_i}))$
     $a_{n_i,j} = (a_{n_i,j} - \min(A_{n_i}))/(\max(A_{n_i})\text{-}\min(A_{n_i}))$
   else
     $a_{n_i,j} = \begin{cases} 1 & \text{if } n_i = j \\ 0 & \text{otherwise} \end{cases}$
4. Define a matrix $P = \{p_{i,j}\}$ $(1 \leq i \leq T, 1 \leq j \leq N)$
   For $j = 1$ to $N$
     For $i = 1$ to $T$
       $p_{i,j} = a'_{n_i} \times a_{n_i,j}$
     end
     $v_j$ = the maximal value of the $j^{th}$ column of $P$
   end
5. Normalize the vector $V = \{v_j\}$ $(1 \leq j \leq N)$ as:
   $v_j = \frac{v_j}{\sum_{1 \leq i \leq N} v_i}$
```

N is the total number of images in the database, and the features of the query image q are denoted as $\{o_1, o_2, ..., o_T\}$, where T is the total number of non-zero features of the query image q. In this study, $1 \leq T \leq 19$ since there are 19 features in total.

Definition 4.3. $W_t(i)$ is defined as the edge weight from image i to q at the evaluation of the t^{th} feature (o_t) in the query, where $1 \leq i \leq N$ and $1 \leq t \leq T$.

Based on the definition, the retrieval algorithm is given as follows. At $t = 1$,

$$W_1(i) = (1 - |b_i(o_1) - b_q(o_1)|/b_q(o_1)) \tag{4.4}$$

The value of $W_{t+1}(i)$, where $1 \leq t \leq T - 1$, is calculated by using the value of $W_t(i)$.

$$W_{t+1}(i) = W_t(i) a_{q,i} (1 - |b_i(o_{t+1}) - b_q(o_{t+1})|/b_q(o_{t+1})) \tag{4.5}$$

As mentioned before, $\mathcal{A}$ represents the relative affinity measures of the semantic relationships among the images in the probabilistic semantic network and $\mathcal{B}$ contains the

Table 4.4: Image retrieval steps using the proposed framework.

1. Given the query image q, obtain its feature vector $\{o_1, o_2, ..., o_T\}$, where T is the total number of non-zero features of the query image q.
2. Upon the first feature o_1, calculate $W_1(i)$ according to Eq. (4.4).
3. To generate the initial query results, set $a_{q,i}$ to be the value of $(q,i)^{th}$ entry in matrix $\mathcal{A}$. Otherwise, based on the user feedback, calculate vector V_q by using the algorithm presented in Table 4.3 and let $a_{q,i}$ equal to v_i, the i^{th} entry in V_q.
4. Move on to calculate $W_2(i)$ according to Eq. (4.5).
5. Continue to calculate the next values for the W vector until all the features in the query have been taken care of.
6. Upon each non-zero feature in the query image, a vector $W_t(i)$ $(1 \leq t \leq T)$ can be obtained. Then each value at the same position in the vectors $W_1(i)$, $W_2(i)$, ..., $W_T(i)$ is summed up. Namely, $sumW_T(i) = \sum_T W_t(i)$ is calculated.
7. Find the candidate images by sorting their corresponding values in $sumW_T(i)$. The bigger the value is, the stronger the relationship that exists between the candidate image and the query image.

low-level features. To generate the initial query results, the value of $a_{q,i}$ from matrix $\mathcal{A}$ is used. Once the user provides the feedback, a vector V_q is calculated by using the algorithm presented in Table 4.3. Then $a_{q,i} = v_i$ $(v_i \in V_q)$ is applied in Eq. (4.5). The stochastic process for image retrieval by using the dynamic programming algorithm is shown in Table 4.4.

Experiments

In the above section, a framework is presented where the semantic network and low-level features can be integrated seamlessly into the image retrieval process to improve the query results. In this section, the experimental results are presented to demonstrate the effectiveness of this framework.

In the experiments, 10,000 color images from the Corel image library with more than 70 categories, such as people, animal, etc., are used. In order to avoid bias and to capture the general users' perceptions, the training process was performed by a group of 10 university students, who were not involved in the design and development of the framework and have no knowledge of the image content in the database. Currently,

Table 4.5: The category distribution of the query image set.

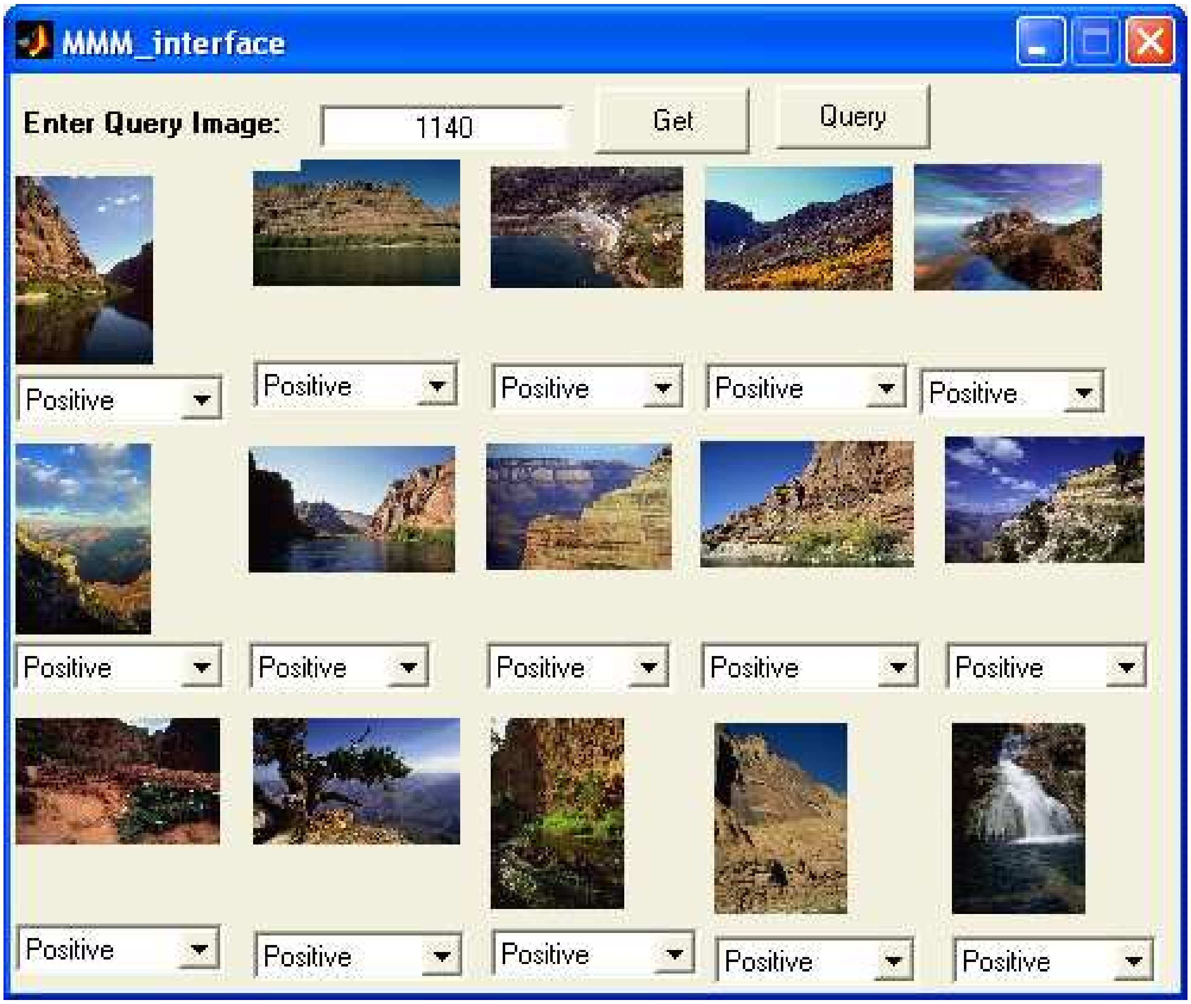
Category	**Explanation**	**Number of Query Images**
Landscape	Land, Sky, Mountain	16
Flower	Flower	16
Animal	Elephant, Panther, Tiger	16
Vehicle	Car, Bus, Plane, Ship	16
Human	Human	16

1,400 user access patterns have been collected through the training system, which covered less than half of the images in the database. The $\mathcal{A}$ matrix and the semantic network are constructed according to the algorithms presented earlier. For the low-level image features, the color and texture features of the images are considered and the $\mathcal{B}$ matrix is obtained by using the procedures illustrated. The constructions of these matrices can be performed off-line.

To test the retrieval performance and efficiency of the proposed mechanism, 80 randomly chosen images belonging to 5 distinct categories were used as the query images. Table 4.5 lists the descriptions for each category as well as the number of query images selected from each category.

For a given query image issued by a user, the proposed stochastic process is conducted to dynamically find the matching images for the user's query. The similarity scores of the images with respect to certain query image are determined by the values in the resulting $sumW_T$ vectors according to the rules described in Table 4.4. Fig. 4.5 gives a query-by-image example, in which the retrieved images are ranked and displayed in the descending order of their similarity scores from the top left to the bottom right, with the upper leftmost image being the query image. In this example, the query image belongs to the 'Landscape' category. As can be seen from this figure, the perceptions contained in these returned images are quite similar and the ranking is reasonably good.

In order to demonstrate the performance improvement and the flexibility of the proposed model, the accuracy-scope curve is used to compare the performance of this mech-



Enter Query Image: 1140

Positive Positive Positive Positive Positive

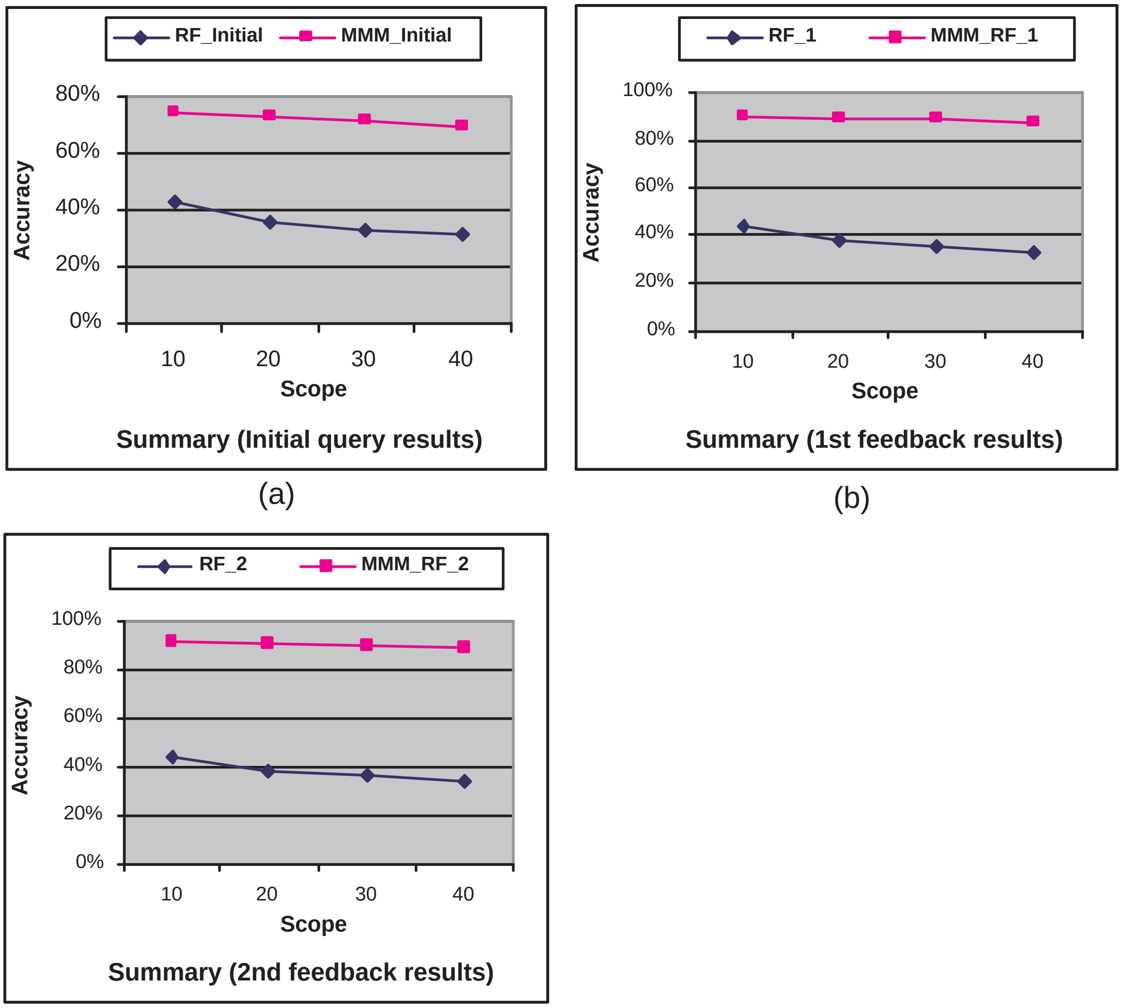
Positive Positive Positive Positive Positive

Positive Positive Positive Positive Positive

Figure 4.5: The snapshot of a query-by-image example.

anism with a common relevance feedback method. In the accuracy-scope curve, the scope specifies the number of images returned to the users and the accuracy is defined as the percentage of the retrieved images that are semantically related to the query image.

In the experiments, the overall performance of the proposed MMM mechanism is compared with the relevance feedback method (RF) proposed in [94] in the absence of the information of user access patterns and access frequencies. The RF method proposed in [94] conducts the query refinement based on re-weighting the low-level image features (matrix $\mathcal{B}$) alone. In fact, any normalized vector-based image feature set can be plugged into the matrix $\mathcal{B}$. Figure 4.6 shows the curves for the average accuracy values



(a)

(b)

(c)

Figure 4.6: Performance comparison.

of the proposed CBIR system and the RF CBIR system, respectively. In Figs. 4.6(a), 'MMM_Initial' and 'RF_initial' indicate the accuracy values of the MMM mechanism and the RF method at the initial retrieval time, respectively. The 'MMM_RF_1(2)' and the 'RF_1(2)' in Figs. 4.6 (b)-(c) represent the accuracy values of the two methods after the first and the second rounds of user relevance feedback. The results in Fig. 4.6 are calculated by using the averages of all the 80 query images. It can be easily observed that this proposed method outperforms the RF method for the various numbers of images retrieved at each iteration. This proves that the use of the user access patterns and access

Table 4.6: Accuracy and efficiency comparison between Relevance Feedback method and the proposed framework.

	Relevance Feedback			**Proposed Framework**		
Category	Number of feedbacks	Feedbacks per image	Accuracy	Number of Feedbacks	Feedbacks per image	Accuracy
Landscape	48	3	55.3%	21	1.3	61.3%
Flower	48	3	44.7%	23	1.4	73.4%
Animal	48	3	48.8%	20	1.3	81.6%
Vehicle	48	3	23.8%	44	2.8	74.4%
Human	48	3	26.9%	33	2.1	75.0%
Summary	240	3	39.9%	141	1.8	73.1%

frequencies obtained from the off-line training process can capture the subjective aspects of the user concepts. As another observation, the proposed method and the RF method share the same trend, which implies that the more the iterations of user feedback, the higher the accuracy they can achieve.

Table 4.6 lists the number of user feedback iterations observed in the RF method and the proposed method for each image category. For example, the number of query images in the ‘Landscape’ category is 16, and the number of user feedback iterations observed for those 16 images is 48 and 21, respectively, for the RF method and the proposed method. Thus, the number of feedback iterations per image is 48/16=3 for the RF method, while it is 1.3 for the proposed method. As can be seen from this table, the proposed method can achieve better retrieval performance even by using a smaller number of feedback iterations than that of the RF method in all five categories.

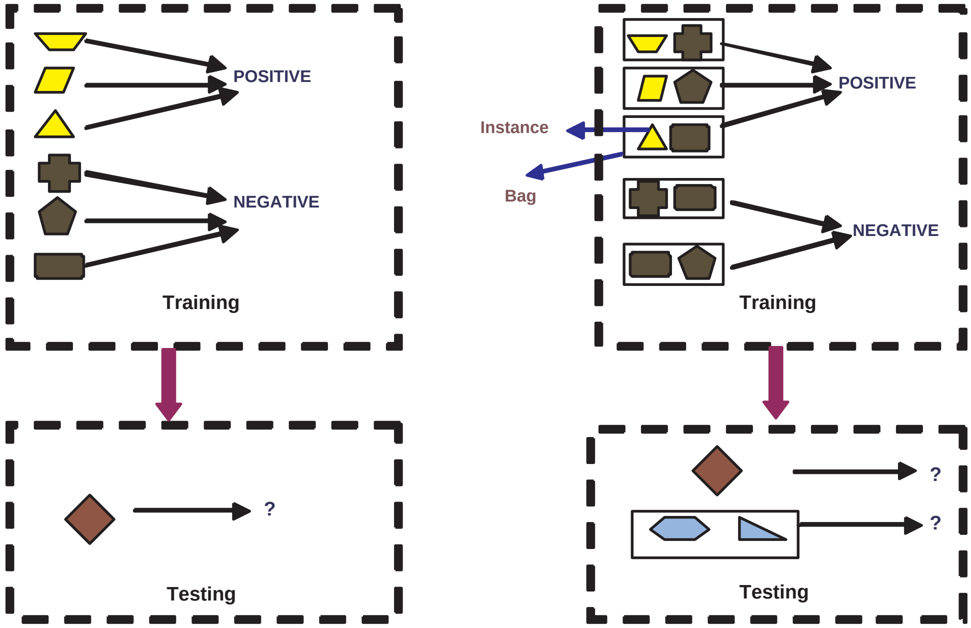
Conclusions

One of the key problems in the CBIR systems come from the concern of lacking a mapping between the high-level concepts and the low-level features. Although Relevance Feedback (RF) has been proposed to address the perception subjectivity issue, the performance is limited by the insufficient power of the low-level features in representing the high-level concepts. In addition, the users are required to take heavy responsibility during the retrieval process to provide feedback in several iterations. The useful information contained in the user feedback is employed to improve the current query results only, without being further utilized to boost the system performance. In response to these issues, a probabilistic semantic network-based image retrieval framework using both relevance feedback and the Markov Model Mediator (MMM) mechanism is proposed. As a result, the semantic network and the low-level features are seamlessly utilized to achieve a higher retrieval accuracy. One of the distinct properties of this framework is that it provides the capability to learn the concepts and affinities among the images, represented by semantic network, off-line based on the training data set, such as access patterns and access frequencies without any user interaction. This off-line learning is in fact an affinity-mining process which can reveal both the inner-query and inter-query image affinities. In addition, the proposed framework also supports the query refinement for individual users in real-time. The experimental results demonstrate the effectiveness and efficiency of this proposed framework for image retrieval.

4.2.2 Hierarchical Learning Framework

As discussed in Chapter 2, by acting alone, the existing CBIR approaches have certain limitations in terms of retrieval accuracy and/or processing costs. In Section 4.2.1, a unified framework is proposed, which integrates the MMM mechanism with the RF technique. However, it intends to bridge the semantic gap and capture the user's perception at the image-level. In this section, the framework is further extended to explore the high-level semantic concepts in a query from both the object-level and the image-level and to address the needs of serving the specific user's query interest as well as reducing the convergence cycles [7].

Specifically, an advanced content-based image retrieval system, MMIR, is proposed [7], where MMM and MIL (the region-based learning approach with Neural Network technique as the core) are integrated seamlessly and act coherently as a hierarchical learning



(a) Idea of traditional supervised learning

(b) Idea of multiple instance learning

Figure 4.7: Overview of the difference between two learning schemes. (a) Idea of traditional supervised learning; (b) Idea of multiple instance learning

engine to boost both the retrieval accuracy and efficiency. By intelligent integration, it aims at offering a potentially promising solution for the CBIR system. As the concept of MMM has been discussed, in the following section, MIL will be introduced first followed by a discussion of the proposed hierarchical learning framework.

Multiple Instance Learning

Motivated by the drug activity prediction problem, Dietterich et al. [30] introduced the Multiple Instance Learning model. Since its introduction, it has become increasingly important in machine learning. The idea of multiple instance learning varies from that of traditional learning problem as illustrated in Fig. 4.7.

As can be seen from Fig. 4.7(a), in a traditional supervised learning problem, the task is to learn a function

$$y = f(x_1, x_2, ..., x_n) \tag{4.6}$$

given a group of examples $(y_i, x_{i1}, x_{i2}, ..., x_{in})$, $i = 1, 2, ..., Z$.

Here, Z represents the number of input examples and n denotes the number of features for each example object. Each set of input values $(x_{i1}, x_{i2}, ..., x_{in})$ is tagged with the label y_i, and the task is to learn a hypothesis (function *f*) that can accurately predict the labels for the unseen objects.

In MIL, however, the input vector $(x_{i1}, x_{i2}, ..., x_{in})$ (called an instance) is not individually labeled with its corresponding y_i value. Instead, one or more instances are grouped together to form a bag $B_b \in \beta$ and are collectively labeled with a $Y_b \in L$, as illustrated in Fig. 4.7(b). The purpose of MIL is that given a training set of bags as well as their labels, it can deduce the label for each instance. Furthermore, since a bag consists of a set of instances, the label of a given bag can be in turn determined. The input/training set of MIL is not as complete as traditional Learning. Here, β denotes the bag space and L represents the label space with $L = \{0(Negative), 1(Positive)\}$ for binary classification. Let α be the instance space and assume there are m instances in B_b, the relation between the bag label Y_b and the labels $\{y_{bj}|y_{bj} \in L\}$ $(j = 1, 2, ..., m)$ of all its instances $\{I_{bj}|I_{bj} \in \alpha\}$ is defined as follows.

$$Y_b = \begin{cases} 1 & \text{if } \exists_{j=1}^{m} y_{bj} = 1 \\ 0 & \text{if } \forall_{j=1}^{m} y_{bj} = 0. \end{cases} \tag{4.7}$$

The label of a bag (i.e., Y_b) is a disjunction of the labels of the instances in the bag (i.e., Y_{bj} where $j = 1, 2, ..., m$). The bag is labeled as positive if and only if at least one of its instances is positive; whereas it is negative when all the instances in that bag are negative. The goal of the learner is to generate a hypothesis $h : \beta \rightarrow L$ to accurately predict the label of a previously unseen bag.

In terms of image representations in the region-based retrieval, images are first segmented into regions, where each of them is roughly homogeneous in color and texture and characterized by a feature vector. Consequently, each image is represented by a collection of feature vectors. From the perspective of learning, the labels (positive or negative) are directly associated with images instead of individual regions. It is reasonable to assume that if an image is labeled as positive, at least one of its regions is of user interest. Intuitively, the basic idea is essentially identical to the MIL settings, where a *bag* refers to an *image*; whereas an *instance* corresponds to a *region*. With the facilitation of MIL, it can be expected to have a reasonably good query performance by discovering and applying the query-related objects in the process and filtering out the irrelevant objects.

In this study, for the sake of accuracy, the real-valued MIL approach developed in our earlier work [51] is adopted. The idea is to transfer the discrete label space $L = \{0(Negative), 1(Positive)\}$ to a continuous label space $L_R = [0, 1]$, where the value indicates the degree of positive for a bag, with label '1' being 100% positive. Therefore, the goal of the learner is to generate a hypothesis $h_R : \beta \rightarrow L_R$. Consequently, the label of the bag (i.e., the degree of the bag being positive) can be represented by the maximum of the labels of all its instances and Eq. 4.7 is then transformed as follows.

$$Y_b = max_j\{y_{bj}\} \tag{4.8}$$

Let $h_I : \alpha \rightarrow L_R$ be the hypothesis to predict the label of an instance, the relationship between hypotheses h_R and h_I is depicted in Eq. 4.9.

$$Y_b = h_R(B_b) = max_j\{y_{bj}\} = max_j\{h_I(I_{bj})\} \tag{4.9}$$

Then the Minimum Square Error (MSE) criterion is used. That is, it tries to learn the hypotheses $\overline{h}_R$ and $\overline{h}_I$ to minimize the following function.

$$S = \sum_b (Y_b - \overline{h}_R(B_b))^2 = \sum_b (Y_b - max_j \overline{h}_I(I_{bj}))^2. \tag{4.10}$$

In this study, the Multilayer Feed-Forward Neural Network is adopted to represent the hypothesis $\overline{h}_I$ and the back-propagation learning method is used to train the neural network to minimize S. More detailed discussion can be found in [51]. In the Experimental Section, the structure and parameter settings of the neural network are discussed. To some extent, the MIL approach can be considered as a hybrid of the RF technique and the region-based retrieval. MIL intends to achieve better query results in the next round by analyzing the training bag labels (i.e., user's feedback), which resembles the RF concepts.

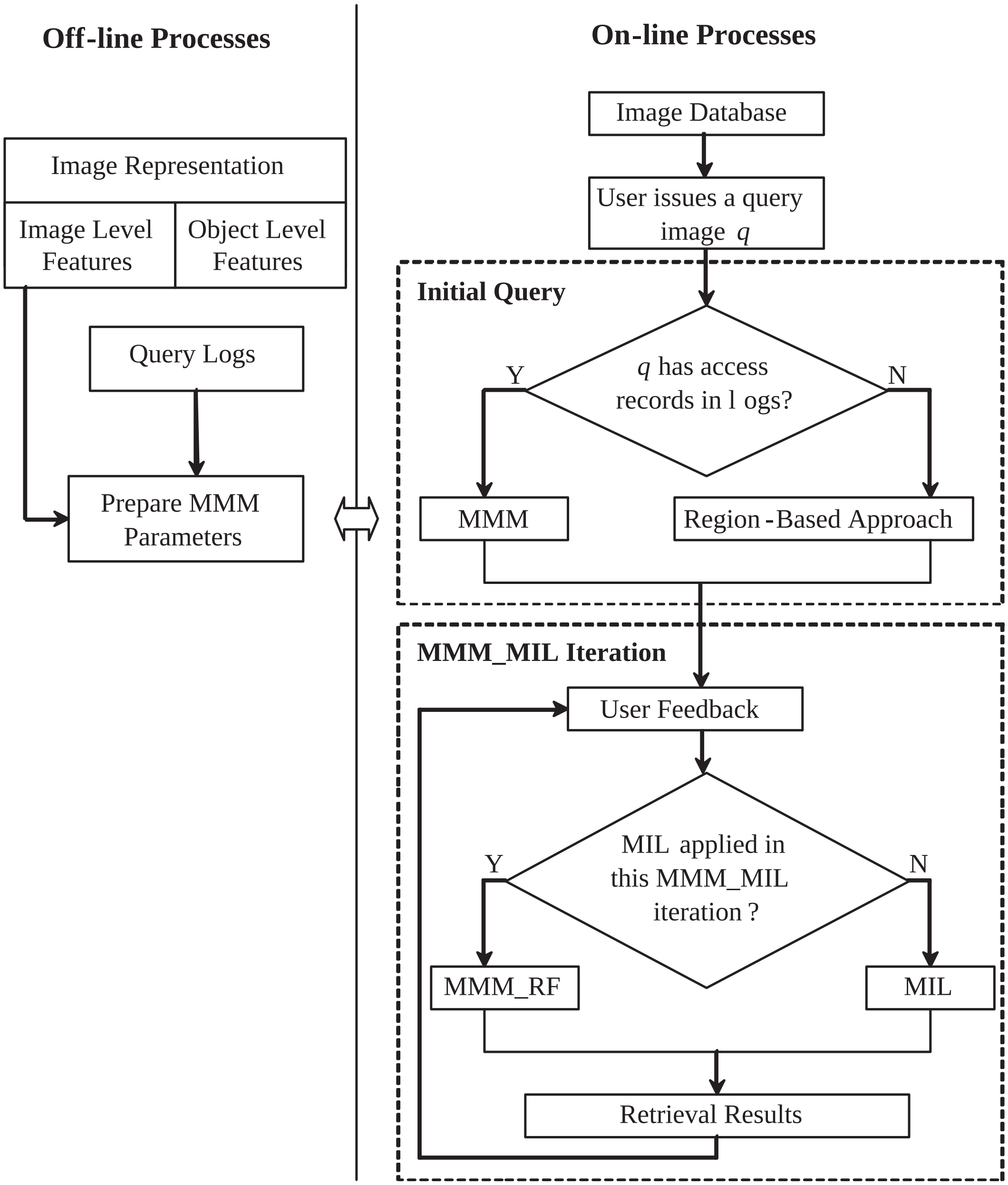


Figure 4.8: The Hierarchical Learning Framework.

Nevertheless, the main focus of MIL is to explore the region of users' interest, which is the reason that MIL can be classified as a region-based approach.

Hierarchical Learning Scheme

As discussed earlier, integrating the essential functionalities from both MMM and MIL has potential in constructing a robust CBIR.

In this subsection, the basic idea and procedure of constructing the hierarchical learning framework (for short, MMM_MIL framework) is presented by integrating these two techniques for the MMIR system, which is illustrated in Fig. 4.8. As can be seen in this figure, the MMM_MIL framework consists of an off-line process which aims at extracting the image and object-level features to obtain the MMM parameters, and an on-line retrieval process. These two processes work closely with each other in the sense that the off-line process prepares the essential data for the on-line process to reduce the on-line processing time. In addition, the feedback provided in the on-line process can be accumulated in the logs for the off-line process to update the MMM parameters periodically. In this section, the focus is on the on-line retrieval process.

- Initial Query

  In most of the existing CBIR systems, given a query image, the initial query results are simply computed by using a certain similarity function (e.g., Euclidean distance, Manhattan distance, etc.) upon the low-level features either in the image or the object level. For instance, in the general MIL framework, since there is no training data available for the outset of the retrieval process, a simple distance-based metric is applied to measure the similarity of two images [51]. Formally, given a query image q with R_q regions (denoted as $q = \{q_i\}, i = 1, 2, ..., R_q$), its difference with respect to an image m consisting of R_m regions (denoted as $m = \{m_j\}, j = 1, 2, ..., R_m$) is defined as:

$$Dist(q,m) = \sum_i min_j\{|q_i - m_j|\}. \qquad (4.11)$$

Here, $|q_i - m_j|$ represents the distance between two feature vectors of regions q_i and m_j. However, due to the semantic gap issue, it is highly possible that the number of "positive" images retrieved in the initial run is relatively small (e.g., less than 5 positives out of the top 30 images). This lack of positive samples greatly hinders the learning performance for most of the learning algorithms, including the NN-based MIL approach discussed earlier. In contrast, MMM possesses the capability of representing the general concepts in the query and outperforms the region-based approach defined in Eq. 4.11 on the average. One exception, though, is that any query image that has not been accessed before will force the MMM mechanism to perform a similarity match upon the low-level image features as discussed in Section 4.2.1. In this case, the region-based approach will be applied as it captures more completed information. Therefore, in the proposed hierarchical learning framework, the initial query is carried out as illustrated in Fig. 4.8. It is worth noting that the test of whether an image q has been accessed before (its access record) in the log can be formally transformed to test whether $\sum_j a(q,j)$ equals 0, where $a(q,j) \in \mathcal{A}$.

- MMM_MIL iteration

  With the initial query results, the users are asked to provide the feedback for the MMM_MIL iteration, which is defined as an MIL process followed by MMM. The basic idea is that based on the region of interest (e.g., instance I_p in image or bag B_p) MIL learned for a specific user, the semantic network represented by MMM is intelligently traversed to explore the images which are semantically related to B_p. Obviously, it can be easily carried out by treating B_p as the query image and using the algorithms described in Section 4.2.1.

Specifically, if a group of positive bags (images) are identified, which is actually the general case, the situation becomes relatively complicated in the sense that a number of paths need to be traversed and the results are then aggregated to reach the final outputs. Therefore, the extended MMM mechanism, MMM_RF, is used to solve this problem. The difference between MMM and MMM_RF is that MMM considers only the direct relationship between the query image q and the other images in the database; whereas MMM_RF adopts an additional relationship called Indirectly related (R_I) relationship which denotes the situation when two images are connected to a common image. With the introduction of R_I, the multiple paths mentioned above can be effectively merged into a new path, where the same dynamic programming based stochastic output process can be applied to produce the final results (please refer to Section 4.2.1).

Experiments

To perform rigorous evaluation of the proposed framework, 9,800 real-world images from the Corel image CDs were used, where every 100 images represent one distinct topic of interest. Therefore, the data set contains 98 thematically diverse image categories, including antique, balloon, car, cat, firework, flower, horse, etc., where all the images are in JPEG format with size 384*256 or 256*384.

In order to evaluate the performance of the proposed MMIR system, the off-line process needs to be carried out first, which includes feature extraction and query log collection. In addition, the neural network structure for MIL should be defined before the on-line process can be conducted.

- **Image Representation.**

  Each image is represented by the color and texture features extracted from both the image and object levels as discussed in Section 4.1.1 and 4.1.2, respectively.

- **Query Logs.**

  The collection of query logs is a critical process for learning the essential parameters in this framework. Therefore, in MMIR, a group of 7 users were asked to create the log information. The users were requested to perform query-by-example (QBE) execution on the system and provide their feedback on the retrieved results.

  In order to ensure that the logs cover a wide range of images, each time a query image is randomly seeded from the image database and the system returns the top 30 ranked images by employing the region-based approach defined earlier. The user then provides the feedback (positive or negative) on the images by judging whether they are relevant to the query image. Such information is named as a query log and is accumulated in the database. Currently, 896 query logs have been collected. Though the users may give noisy information to the logs, it will not significantly affect the learning performance as long as it only accounts for a small portion of the query logs.

- **Neural Network.**

  As discussed earlier, a three-layer Feed-Forward Neural Network is used to map an image region with a low-level feature vector into the user's high-level concept.

  As can be seen from Fig. 4.9, the network consists of an input layer, a hidden layer and an output layer. Here, the input layer contains 19 input units, where each of them represents a low-level feature of an image region. Therefore, the notations $f_1, f_2, ..., f_{19}$ correspond to the 19 low-level features described previously. The hidden layer is composed of 19 hidden nodes with w_{ij} being the weight of the connection between the i^{th} input unit I_i and the j^{th} hidden node H_j (where $i, j = 1, 2, ..., 19$). The output layer contains only one node, which outputs the real value $y \in L_R = [0, 1]$ indicating the satisfactory level of an image region with regard to a user's concept. The weight between the output node and the j_{th} hidden

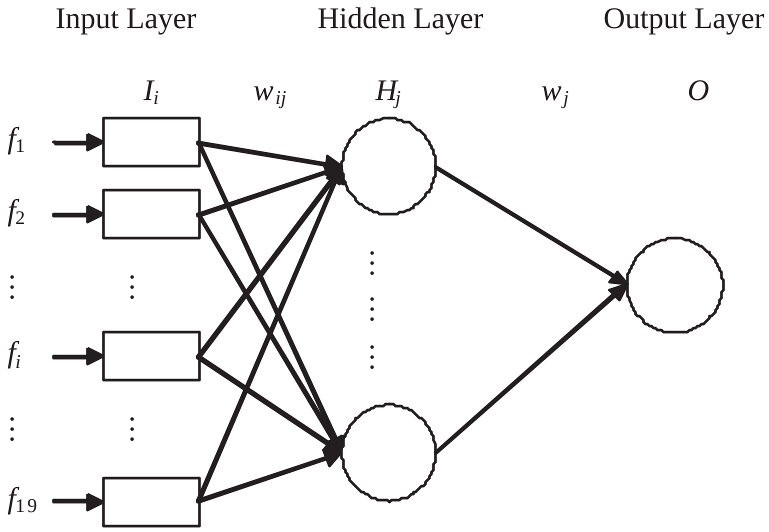


Figure 4.9: The three-layer Feed-Forward Neural Network.

node H_j is in turn denoted as w_j. The Sigmoid function with slope parameter 1 is used as the activation function and the back-propagation (BP) learning method is applied with a learning rate of 0.1 with no momentum. The initial weights for all the connections (i.e., w_{ij} and w_j) are randomly set with relatively small values (e.g., in the range of [-0.1, 0.1]) and the termination condition of the BP algorithm is defined as follows.

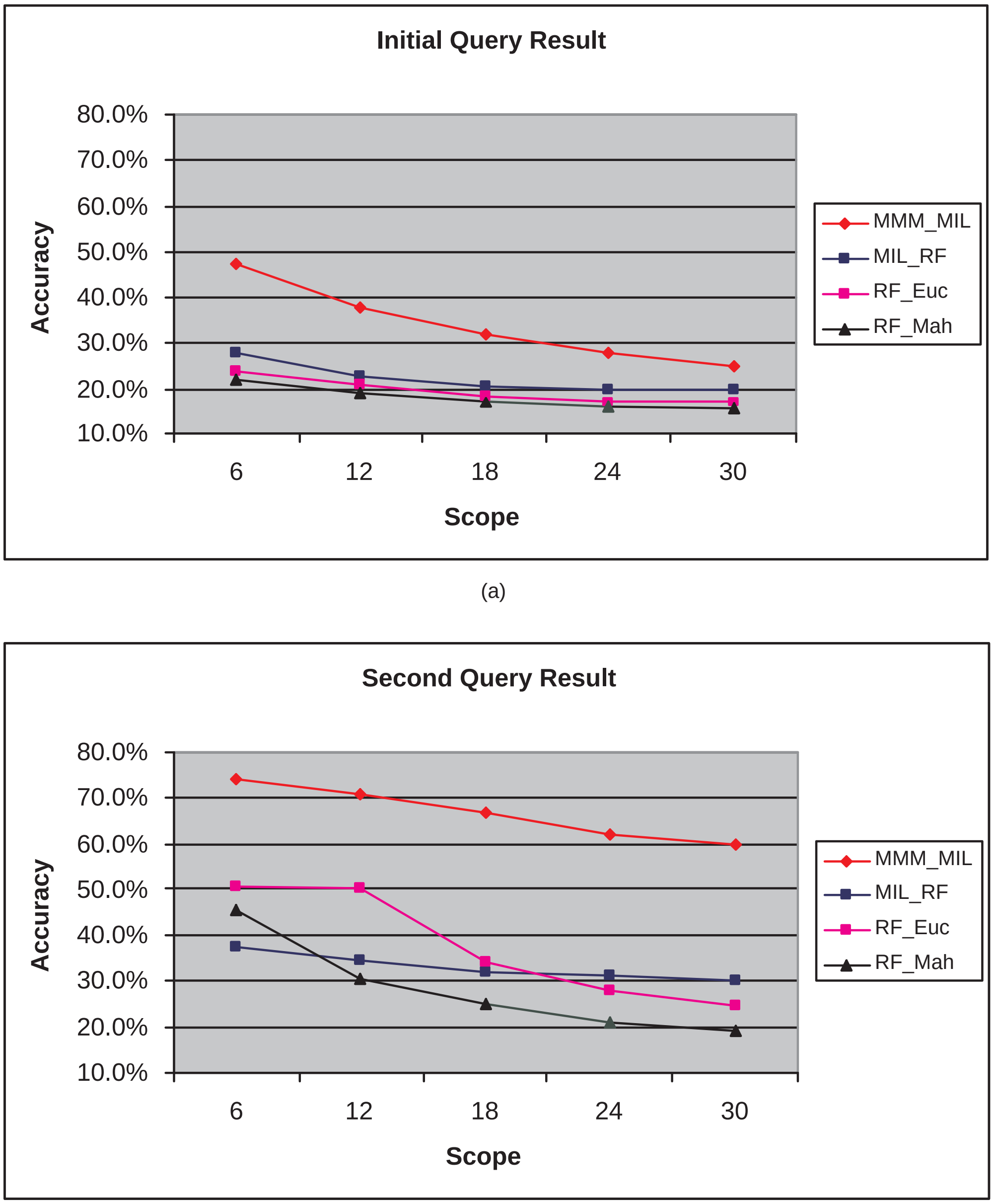
$$|S^{(k)} - S^{(k-1)}| < \alpha \times S^{(k-1)}. \tag{4.12}$$

Here, $S^{(k)}$ denotes the value of S at the k_{th} iteration and α is a small constant, which is set to 0.005 in the experiment.

As usual, the performance measurement metric employed in the experiments is accuracy, which is defined as the average ratio of the number of relevant images retrieved over the number of total returned images (or called scope).

In order to evaluate the performance of the hierarchical learning framework (denoted as MMM_MIL), it is compared with the Neural Network based MIL technique with relevance feedback (for short, MIL_RF) which does not support the log-based retrieval.

In addition, its performance is also compared with another general feature re-weighting algorithm [94] with relevance feedback using both Euclidean and Manhattan distances, denoted as RF_Euc and RF_Mah, respectively.



(a)

(b)

Figure 4.10: MMIR Experimental Results.

Fifty query images are randomly issued. For each query image, the Initial query results are first retrieved and then followed by two rounds of user feedback with regard to MIL_RF, RF_Euc and RF_Mah algorithms. Correspondingly, besides the initial query, one MMM_MIL iteration is performed as each iteration consists of two rounds of feedback. In the database log, totally 896 distinct queries have been recorded which are used by MMM_MIL. In addition, the region-level features used by MIL_RF are the same as the ones used by MMM_MIL. Similarly, the image-level features used by RF_Euc, RF_Mah and MMM_MIL are also identical.

The accuracy within different scopes, i.e., the percentages of positive images within the top 6, 12, 18, 24, and 30 retrieved images are calculated. The results are illustrated in Fig. 4.10, where Figs. 4.10(a) and 4.10(b) show the initial query results and the second query (or the first round of MMM_MIL) results, respectively.

As can be seen from this figure, the accuracy of MMM_MIL greatly outperforms all the other three algorithms in all the cases. More specifically, with regard to the initial query results (Fig. 4.10(a)), MMM_MIL (represented by the red line) performs far better than the remaining three algorithms with more than 10% difference in accuracy on average, which demonstrates MMM's strong capability in capturing the general concepts. Furthermore, by comparing Fig. 4.10(a) and Fig. 4.10(b), it can be observed that the MMM_MIL results improve tremendously where the increment of the accuracy rate reaches 30% on average. In contrast, the improvements of the other approaches are relatively small (with the improvement of the accuracy rate ranging from 10% to 20%), which indicates that MMM_MIL can achieve an extremely fast convergence of the concept.

Conclusions

As an emerging topic, the application of the learning techniques in the CBIR system has attracted increasing attention nowadays. With the aim of addressing the semantic gap and the perception subjectivity issues, an advanced content-based image retrieval system called MMIR is proposed in this section that is facilitated with a hierarchical learning framework called MMM_MIL. The proposed MMIR system utilizes the MMM mechanism to direct the focus on the image level analysis together with MIL technique (with the Neural Network technique as its core) to real-time capture and learn the object-level semantic concepts with some help of the user feedback. In addition, from a long-term learning perspective, the user feedback logs are explored by MMM to speed up the learning process and to increase the retrieval accuracy. As a result, the unique characteristic of the proposed framework is that it not only possesses strong capabilities in real-time capturing and learning of the object and image semantic concepts, but also offers an effective solution to speed up the learning process. Comparative experiments with the well-known learning techniques fully demonstrate the effectiveness of the proposed MMIR system.

4.2.3 Inter-database Retrieval

The above-discussed approaches are mainly conducted on a single database level, which is not sufficient to meet the increasing demand of handling efficient image database retrieval in a distributed environment. In addition, in the traditional database research area, data clustering places related or similar valued records or objects in the same page on disks for performance purposes. However, due to the autonomous nature of each image database, it is not realistic to improve the performance of databases by actually moving around the databases.

In response to these issues, the MMM mechanism is further extended to enable image database clustering and cluster-based image retrieval for efficiency purposes [106]. In particular, the work is proposed to use MMMs for the construction of probabilistic networks via the affinity mining process, to facilitate the conceptual database clustering and the image retrieval process at both intra-database and inter-database levels. It is a unified framework in the sense that the same mechanism (MMM) is utilized at different hierarchies (local image databases and image database clusters) to build probabilistic networks which represent the affinity relations among images and databases. The proposed database clustering strategy fully utilizes the information contained in the integrated probabilistic networks, and partitions the image databases into a set of conceptual image database clusters without physically moving them. Essentially, since a set of image databases with close relationships are put in the same image database cluster and are required consecutively on some query access path, the number of platter (cluster) switches for data retrieval with respect to the queries can be reduced.

The core of the proposed framework is the MMM mechanism that facilitates conceptual database clustering to improve the retrieval accuracy. An MMM-based conceptual clustering strategy consists of two major steps: 1) calculating the similarity measures between every two image databases, and 2) clustering databases using the similarity measures. Here, two image databases are said to be related if they are accessed together frequently or contain similar images. In the first step, a local probabilistic network is built to represent the affinity relationships among all the images within each database, which is modeled by a local MMM and enables accurate image retrieval at the intra-database level, which has been discussed above and will not be covered in this section. The second step is the proposed conceptual clustering strategy that fully utilizes the parameters of the local MMMs to avoid the extra cost of information summarization, which may be unavoidable in other clustering methods. In our previous work [106], a thorough comparative study has been conducted, in which the MMM mechanism was compared with several clustering algorithms including single-link, complete-link, group-average-link, etc. The experimental results demonstrated that MMMs produce the best performance in general-purpose database clustering. However, it cannot be directly applied to image database clustering because: 1) image data have special characteristics that are quite different from numerical/textual data; and 2) image database queries are different from traditional database queries in that they may involve users subjective perceptions in the retrieval process.

In this study, the general MMM-based clustering strategy is further extended to handle image database clustering. For each image database cluster, an inter-database level probabilistic network, represented by an integrated MMM, is constructed to model a set of autonomous and interconnected image databases in it, which serves to reduce the cost of retrieving images across the image databases and to facilitate accurate image retrieval within the cluster.

Calculating the Similarity Measures

The conceptual image database clustering strategy is to group related image databases in the same cluster such that the intra-cluster similarity is high and the inter-cluster similarity is low. Thus, a similarity measure needs to be calculated for each pair of image databases in the distributed database system. These similarity measures indicate the relationships among the image databases and are then used to partition the databases into clusters.

Let d_i and d_j be two image databases, and $X = \{x_1, ..., x_{k1}\}$ and $Y = \{y_1, ..., y_{k2}\}$ be the set of images in d_i and d_j, where $k1$ and $k2$ are the numbers of the images in X and Y, respectively. Let $N_k = k1 + k2$ and $O^k = \{o_1, ..., o_{Nk}\}$ be an observation set with the features belonging to d_i and d_j and generated by query q_k, where the features $o_1, ..., o_{k1}$ belong to d_i and $o_{k1+1}, ..., o_{Nk}$ belong to d_j. Assume that the observation set O^k is conditionally independent given X and Y, and the sets $X \in d_i$ and $Y \in d_j$ are conditionally independent given d_i and d_j. The similarity measure $S(d_i, d_j)$ is defined in following equation.

$$S(d_i, d_j) = (\sum_{O^i \subset OS} P(O^k|X, Y; d_i, d_j) P(X, Y; d_i, d_j)) F(N_k) \tag{4.13}$$

where $P(X, Y; d_i, d_j)$ is the joint probability of $X \in d_i$ and $Y \in d_j$, and $P(O^k|X, Y; d_i, d_j)$ is the probability of occurrence of O^k given X in d_i and Y in d_j. They are in turn defined as follows:

$$P(O^k|X, Y; d_i, d_j) = \Pi_{u=1}^{k1} P(o_u|x_u) \Pi_{v=k1+1}^{N_k} P(o_v|y_{v-k1}) \tag{4.14}$$

$$P(X, Y; d_i, d_j) = \Pi_{u=2}^{k1} P(x_u|x_{u-1}) \Pi_{v=k1+2}^{N_k} P(y_{v-k1}|y_{v-k1-1}) P(y_1) \tag{4.15}$$

In Eq. 4.14, $P(o_u|x_u)$ (or $P(o_v|y_{v-k1})$) represents the probability of observing a feature o_u (or o_v) from an image x_u (or y_{v-k1}), which as discussed earlier is captured in matrix $\mathcal{B}$ of an individual database. In Eq. 4.15, $P(x_u|x_{u-1})$ (or $P(y_{v-k1}|y_{v-k1-1})$ indicates the probability of retrieving an image x_u (or y_{v-k1}) given the current query image as x_{u-1} (or y_{v-k1-1}), which is represented by the semantic network as introduced above. $P(x_1)$ (or $P(y_1)$) is the initial probability contained in Π, which is called the initial state probability distribution and indicates the probability that an image can be the query image for the incoming queries and is defined as follows.

$$\Pi = \{\pi_m\} = \sum_{k=1}^{q} use_{m,k} / \sum_{l=1}^{N} \sum_{k=1}^{q} use_{l,k} \tag{4.16}$$

Here, N is the number of images in an image database d_i and parameter use is access pattern as defined in Section 4.1.3. Therefore, the similarity values can be computed for each pair of image databases based on the MMM parameters of each individual database (for short, local MMMs). Then a probabilistic network is built with each image database represented as a node in it. For nodes d_i and d_j in this probabilistic network, the branch probability $P_{i,j}$ is transformed from the similarity value $S(d_i, d_j)$. Here, the transformation is performed by normalizing the similarity values per row to indicate the branch probabilities from a specific node to all its accessible nodes.

Clustering Image Databases

Based on the probability distributions for the local MMMs and the probabilities $P_{i,j}$ for the probabilistic network, the stationary probability ϕ_i for each node i of the probabilistic network is computed from $P_{i,j}$, which denotes the relative frequency of accessing node i (the i^{th} image database, or d_i) in the long run.

$\sum_i \phi_i = 1$

$\phi_j = \sum_i \phi_i P_{i,j}, j = 1, 2, ...$

The conceptual image database clustering strategy is traversal based and greedy. Conceptual image database clusters are created according to the order of the stationary probabilities of the image databases. The image database that has the largest stationary probability is selected to start a new image database cluster. While there is room in the current cluster, all image databases accessible in the probabilistic network from the current member image databases of the cluster are considered. The image database with the next largest stationary probability is selected and the process continues until the cluster fills up. At this point, the next un-partitioned image database from the sorted list starts a new image database cluster, and the whole process is repeated until no un-partitioned image databases remain. The time complexity for this conceptual database clustering strategy is $O(plogp)$ while the cost of calculating the similarity matrix is $O(p^2)$, where p is the number of image databases. The size of each image database cluster is predefined and is the same for all image database clusters.

Construction of the Integrated MMMs

As discussed earlier, each image database is modeled by a local MMM. Another level of MMMs (called integrated MMMs) is also constructed in the proposed framework, which is used to represent the conceptual image database cluster to model a set of autonomous and interconnected image databases within it and to reduce the cost of retrieving images across image databases and to facilitate accurate image retrieval. The cluster-based image retrieval is then supported by using the integrated MMM.

For any images s and t in a conceptual image database cluster CC, the formulas to calculate $\mathcal{A}$ are defined in Definition 4.5. Here, it is assumed that CC contains two or more image databases; otherwise, $\mathcal{A}$ is calculated the same as the one defined for a single image database.

Definition 4.5. Let λ_i and λ_j denote two local MMMs for image databases d_i and d_j, where $j \neq i$ and $\lambda_i, \lambda_j \in CC$.

- $p_{s,t} = f_{s,t} / \sum_{n \in CC} f_{s,n}$ = the probability that λ_i goes to λ_j with respect to s and t; where $f_{s,t}$ are defined similarly as aff_m, n in Definition 4.2, except that they are calculated in CC instead of a single image database;

- $p_s = 1 - \sum_{t \notin \lambda_i} p_{s,t}$ = the probability that λ_i stays with respect to s;

- $a_{s,t}$ = the conditional probability of a local MMM;

- $a^{'}_{s,t}$ = the state transition probability of an integrated MMM, where if $s, t \in \lambda_i \Rightarrow a^{'}_{s,t} = p_s a_{s,t}$, and if $s \in \lambda_i \wedge t \notin \lambda_i \Rightarrow a^{'}_{s,t} = p_{s,t}$;

$\mathcal{A}$ is obtained by repeating the above steps for all local MMMs in CC. As for $\mathcal{B}$ and Π in the integrated MMM, the construction methods are similar to those for local MMM, except that the image scope is defined in the cluster CC.

Once the integrated MMMs are obtained, content-based retrieval can be conducted at the image database cluster level similarly as defined in Definition 4.3 and then the similarity function is defined as:

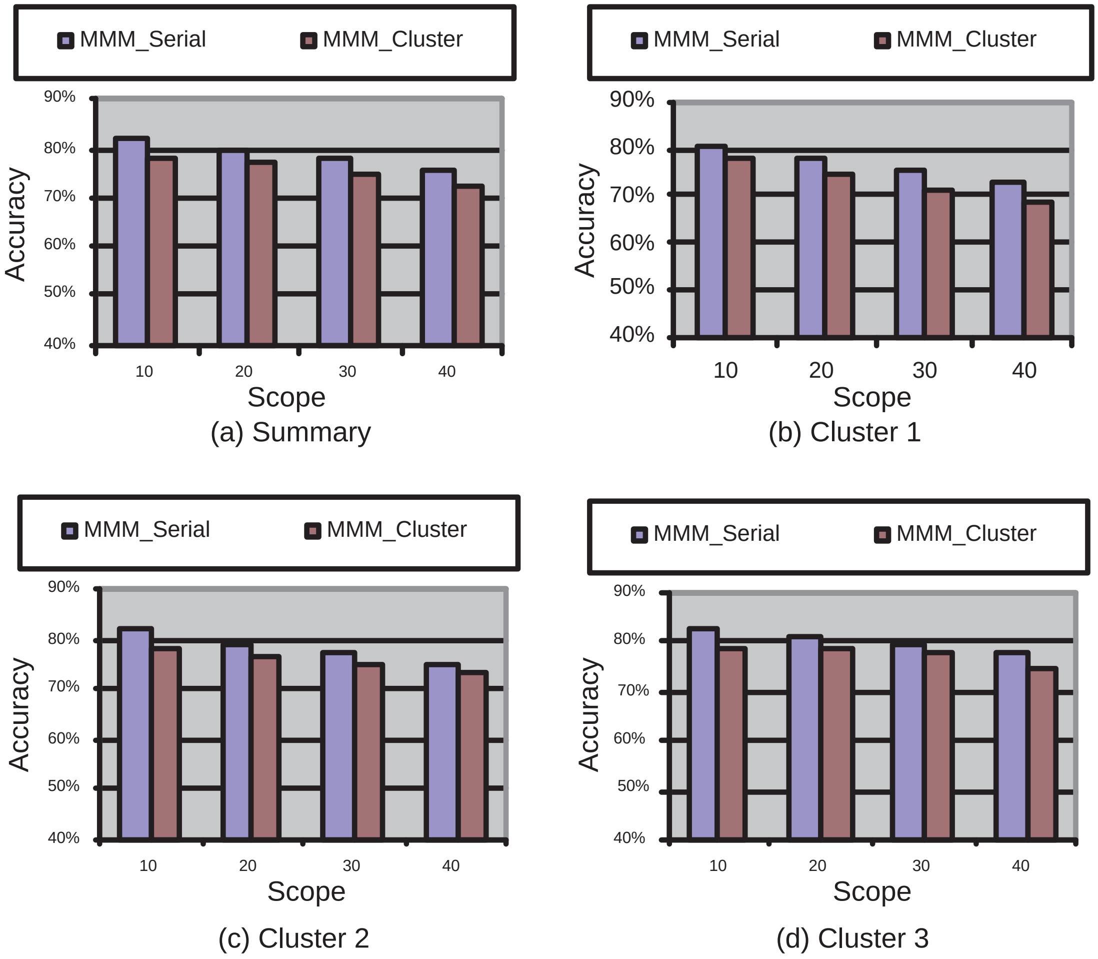
$$SS(q, i) = \sum_{t=1}^{T} W_t(q, i) \tag{4.17}$$

$SS(q, i)$ represents the similarity score between images q and i, where a larger score suggests higher similarity. Note that the same retrieval algorithms can be applied to image retrieval at both local database and database cluster levels by using local or integrated MMMs. Its effectiveness in image retrieval at the local database level have been demonstrated in [99]. In this study, the effectiveness of the proposed framework in conceptual image database clustering and inter-database level image retrieval is examined.

Experimental Results

To show the effectiveness of image retrieval in conceptual image database clusters, 12 image databases with a total of 18,700 images (the number of images in each image database ranges from 1,350 to 2,250) are used. Affinity-based data mining process is conducted utilizing the training data set, which contains the query trace generated by 1,400 queries issued to the image databases. The proposed conceptual image database clustering strategy is employed to partition these 12 image databases into a set of image database clusters. Here, the size of the conceptual image database cluster is set to 4, which represents the maximal number of member image databases a cluster can have. As a result, 3 clusters are generated with 6,450, 5,900 and 6,350 images, respectively. The performance is tested by issuing 160 test queries to these 3 clusters (51, 54 and 55 queries, respectively). For comparison, an image database (namely *DB_whole*) with all the 18,700 images is constructed and tested by the same set of the queries.

Figure 4.11 shows the comparison results, where the scope specifies the number of images returned and the accuracy at a scope s is defined as the ratio of the number of the relevant images within the top s images. In this Figure, ‘*MMM_Cluster* represents the retrieval accuracy achieved by issuing queries to each of the database clusters, while ‘*MMM_Serial* denotes the results of carrying out the search throughout the *DB_whole* image database. For instance, ‘*MMM_Cluster* and ‘*MMM_Serial* in Fig. 4.11(b) represent the results obtained by issuing 51 queries to cluster 1 and *DB_whole*, respectively. As shown in this figure, the accuracy of ‘*MMM_Cluster* is slightly worse than ‘*MMM_Serial*, which is reasonable because ‘*MMM_Cluster* carries out the search in a subspace of *DB_whole*. Considering that the search space is reduced at about one third of the whole space and the image retrieval is conducted at the inter-database level, the effectiveness of the proposed framework in both conceptual image database clustering and content-based image retrieval is obvious. By using the conceptual image database



(a) Summary

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

Figure 4.11: Performance comparison.

clusters, the query cost can be reduced dramatically (almost 1/3) without significant decreases in accuracy (averagely 3%).

Conclusions

In this section, Markov Model Mediators (MMMs), a mathematically sound framework, is extended to facilitate both the conceptual image database clustering and the cluster-based content-based image retrieval. The proposed framework takes into consideration both the efficiency and effectiveness requirements in content-based image retrieval. An effective database clustering strategy is employed in the framework to partition the image databases into a set of conceptual image database clusters, which reduces the query cost dramatically without decreasing the accuracy significantly. In addition, the affinity relations among the images in the databases are explored through the data mining process, which capture the users concepts in the retrieval process and significantly improve the retrieval accuracy.

CHAPTER 5

Data Management for Video Database

With the proliferation of video data, there is a great need for advanced techniques for effective video indexing, summarization, browsing, and retrieval. In terms of modeling and mining videos in a video database system, there are two widely used schemes - shot-based approach and object-based approach. The shot-based approach divides a video sequence into a set of collections of video frames with each collection representing a continuous camera action in time and space, and sharing similar high-level features (e.g., semantic meaning) as well as similar low-level features like color and texture. In the object-based modeling approach, temporal video segments representing the life-span of the objects as well as some other object-level features are used as the basic units for video mining. Object-based modeling is best suitable where a stationary camera is used to capture a scene (e.g., video surveillance applications). However, the video sequences in most of the applications (such as sports video, news, movies, etc.) typically consists of hundreds of shots, with their durations ranging from seconds to minutes.

In addition, the essence of the video is generally represented by important activities or events, which are of users' interests in most cases. This dissertation thus aims to offer a novel approach for event-level indexing and retrieval. Since shot is normally regarded as a self-contained unit, it is reasonable to define the event at the shot-level. Therefore, a shot-based approach is adopted in the proposed framework in terms of video data management. It is worth noting that the state-of-the-art event detection frameworks are generally conduced toward the videos with loose structures or without story units, such as sports videos, surveillance videos, or medical videos [148]. In this chapter, an intelligent shot boundary detection algorithm is briefly introduced followed by the discussions of shot-level video data representation, indexing and retrieval. Similar to the organization of Chapter 4, the focus of this chapter is on the mid-level representation to bridge the

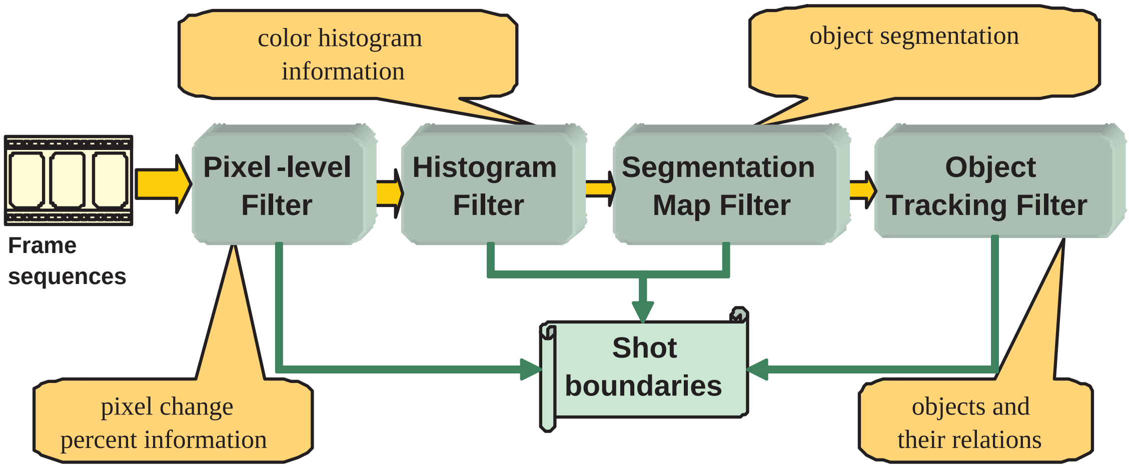


Figure 5.1: The multi-filtering architecture for shot detection.

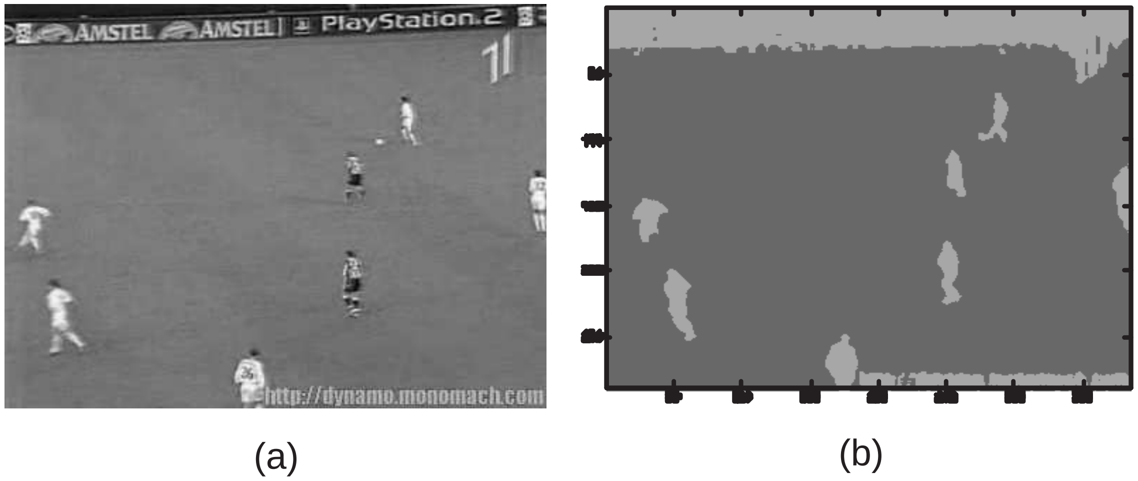
semantic gap and high-level video indexing which is based on the event detection. Due to its popularity, soccer videos are selected as the test bed in this chapter. In addition, this framework can be further extended for concept extraction as will be discussed later, where the concepts refer to high-level semantic features, like “commercial,” “sports,” etc. [117]. The concept extraction schemes are largely carried out on the news videos which have content structures. One of the typical driven forces is the creation of the TRECVID benchmark by National Institute of Standards and Technology, which aims to boost the researches in semantic media analysis by offering a common video corpus and a common evaluation procedure. In addition, an expanded multimedia concept lexicon is being developed by the LSCOM workshop [73] on the order of 1000.

5.1 Video Shot Detection

Video shot change detection is a fundamental operation used in many multimedia applications involving content-based access to video such as digital libraries and video on demand, and it is generally performed prior to all other processes. Although shot detection has a long history of research, it is not a completely solved problem [46], especially for sports videos. According to [32], due to the strong color correlation between soccer shots, a shot change may not be detected since the frame-to-frame color histogram difference is not significant. Secondly, camera motions and object motions are largely present in soccer videos to track the players and the ball, which constitute a major source of false positives in shot detection. Thirdly, the reliable detection of gradual transitions, such as fade in/out, is also needed for sports videos. The requirements of real-time processing need to be taken into consideration as it is essential for building an efficient sports video management system. Thus, a three-level filtering architecture is applied for shot detection, namely pixel-histogram comparison, segmentation map comparison, and object tracking as illustrated in Fig. 5.1. The pixel-level comparison basically computes the differences in the values of the corresponding pixels between two successive frames. This can, in part, solve the strong color-correlation problem because the spatial layout of colors also contributes to the shot detection.

However, though simple as it is, it is very sensitive to object and camera motions. Thus, in order to address the second concern of camera/object motions, a histogram-based comparison is added to pixel-level comparison to reduce its sensitivity to small rotations and slow variations. However, the histogram-based method also has problems. For instance, two successive frames will probably have the similar histograms but with totally different visual contents. On the other hand, it has difficulty in handling the false positives caused by the changes in luminance and contrast.

The reasons of combining the pixel-histogram comparison in the first level filtering are two folds: 1) Histogram comparison can be used to exclude some false positives due to the sensitivity of pixel comparison, while it would not incur much extra computation because both processes can be done in one pass for each video frame. The percentage of changed pixels (denoted as pixel_change) and the histogram difference (denoted as histo_change) between consecutive frames, obtained in pixel level comparison and histogram comparison respectively, are important indications for camera and object motions and can be used to extract higher-level semantics for event mining. 2) Both of them are computationally



(a)

(b)

Figure 5.2: An example segmentation mask map. (a) An example soccer video frame; (b) the segmentation mask map for (a)

simple. By applying a relatively loose threshold, it can be ensured that most of the correct shot boundaries will be included, and in the meanwhile, a much smaller candidate pool of shots is generated at a low cost.

Since the object segmentation and tracking techniques are much less sensitive to luminance change and object motion, the segmentation map comparison and object tracking processes are implemented based on an unsupervised object segmentation and tracking method proposed in our previous work [15][16].

Specifically, the WavSeg segmentation algorithm introduced in Section 4.1.2 for object-level feature extraction in image database can be applied upon the video frame (a still image) for the purpose of segmentation map comparison and object tracking. Based on the frame segmentation result, the segmentation mask map, which contains significant objects or regions of interest, can be extracted from that video frame. In this study, the pixels in each frame are grouped into different classes (for example, 2 classes), corresponding to the foreground objects and background areas. Then two frames can be compared by checking the differences between their segmentation mask maps. An example segmentation mask map is given in Fig. 5.2. The segmentation mask map comparison is especially effective in handling the fade in/out effects with drastic luminance changes and flash light effects [17]. Moreover, in order to better handle the situation of camera panning and tilting, the object tracking technique based on the segmentation results is used as an enhancement to the basic matching process. Since the segmentation results are already available, the computation cost for object tracking is almost trivial compared to the manual template-based object tracking methods. It needs to be pointed out that there is no need to do object segmentation for each pair of consecutive frames. Instead, only the shots in the small candidate pool will be fed into the segmentation process. The performance of segmentation and tracking is further improved by using incremental computation together with parallel computation [144]. The time for segmenting one video frame ranges from 0.03∼0.12 second depending on the size of the video frames and the computer processing power.

In essence, the basic idea for this algorithm is that the simpler but more sensitive checking steps (e.g., pixel-histogram comparison) are first carried out to obtain a candidate pool, which thereafter is refined by the methods that are more effective but with a relatively higher computational cost.

5.2 Video Data Representation

Sports video analysis, especially sports events detection, has received a great deal of attention [28][32][53][139] because of its great commercial potentials. As reviewed in Chapter 2, most existing event detection approaches are carried out in a two-step procedure, that is, to extract the low-level descriptors in a single channel (called unimodal) or multiple channels (called multimodal) and to explore the semantic index from the low-level descriptors using the decision-making algorithm. The unimodal approach utilizes the features of a single modality, such as visual [38][134], audio [133], or textual [140], in soccer highlights detection. For example, [123] proposed a method to detect and localize the goal-mouth in MPEG soccer videos. The algorithm in [67] took advantage of motion descriptors that are directly available in MPEG format video sequences for event detection. In terms of the audio mode, the announcer's excited speech and ball-bat impact sound were detected in [93] for baseball game analysis. For the textual mode, the key words were extracted from the closed captions for detecting events in American football videos [2]. However, because the content of a video is intrinsically multimodal, in which the semantic meanings are conveyed via visual, auditory, and textual channels, such unimodal approaches have their limitations. Currently, the integrated use of multimodal features has become an emerging trend in this area. In [28], a multimodal framework using combined audio, visual, and textual features was proposed. A maximum entropy method was proposed in [44] to integrate image and audio cues to extract highlights from baseball video.

Though multimodal analysis shows promise in capturing more complete information from video data, it remains a big challenge in terms of detecting semantic events from low-level video features due to the well-known semantic gap issue. Intuitively, low-level descriptors alone are generally not sufficient to deliver comprehensive video content. Furthermore, in many applications, the most significant events may happen infrequently, such as suspicious motion events in surveillance videos and goal events in soccer videos. Consequently, the limited amount of training data poses additional difficulties in detecting these so-called rare events. To address these issues, it is indispensable to explore multi-level (low-level, mid-level and high-level) video data representations and intelligently employ mid-level and knowledge-assisted data representation to fill the gap. In this section, the extraction of low-level feature and mid-level descriptors will be introduced. In terms of knowledge-assisted data representation, its main purpose is to largely relax the framework's dependence upon the domain knowledge and human efforts, which is one of the ultimate goals for intelligent data management/retrieval and requires tremendous research efforts. For clarity, Chapter 6 is dedicated to this topic.

In essence, the proposed video event detection framework introduced in this section is shot-based, follows the three-level architecture [31], and proceeds with three steps: low-level descriptor extraction, mid-level descriptor extraction, and high-level analysis. Low-level descriptors, such as generic visual and audio descriptors are directly extracted from the raw video data, which are then used to construct a set of mid-level descriptors including the playfield descriptor (field/grass ratio in soccer games), camera view descriptors (global views, medium views, close-up views, and outfield views), corner view descriptors (wide corner views and narrow corner views), and excitement descriptors. Both of the two modalities (visual and audio) are used to extract multimodal descriptors at low- and mid-level as each modality provides some cues that correlate with the occurrence of video events.

5.2.1 Low-level Multimodal Features

In the proposed framework, multimodal features (visual and audio) are extracted for each shot [9] based on the shot boundary information obtained in the Section 5.1.

Visual Feature Descriptors Extraction

In fact, not only can the proposed video shot detection method detect shot boundaries, but also produce a rich set of visual features associated with each video shot. For examples, the pixel-level comparison can produce the percent of changed pixels between consecutive frames, while the histogram comparison provides the histogram differences between frames, both of which are very important indications for camera and object motions. In addition, the object segmentation can further be analyzed to provide certain region-related information such as foreground/background areas. With these advantages brought by video shot detection, a set of shot-level visual feature descriptors are extracted for soccer video analysis and indexing, namely pixel_change, histo_change, class1_region_mean, class1_region_var, class2_region_mean, and class2_region_var. Here, pixel_change denotes the average percentage of the changed pixels between the consec-

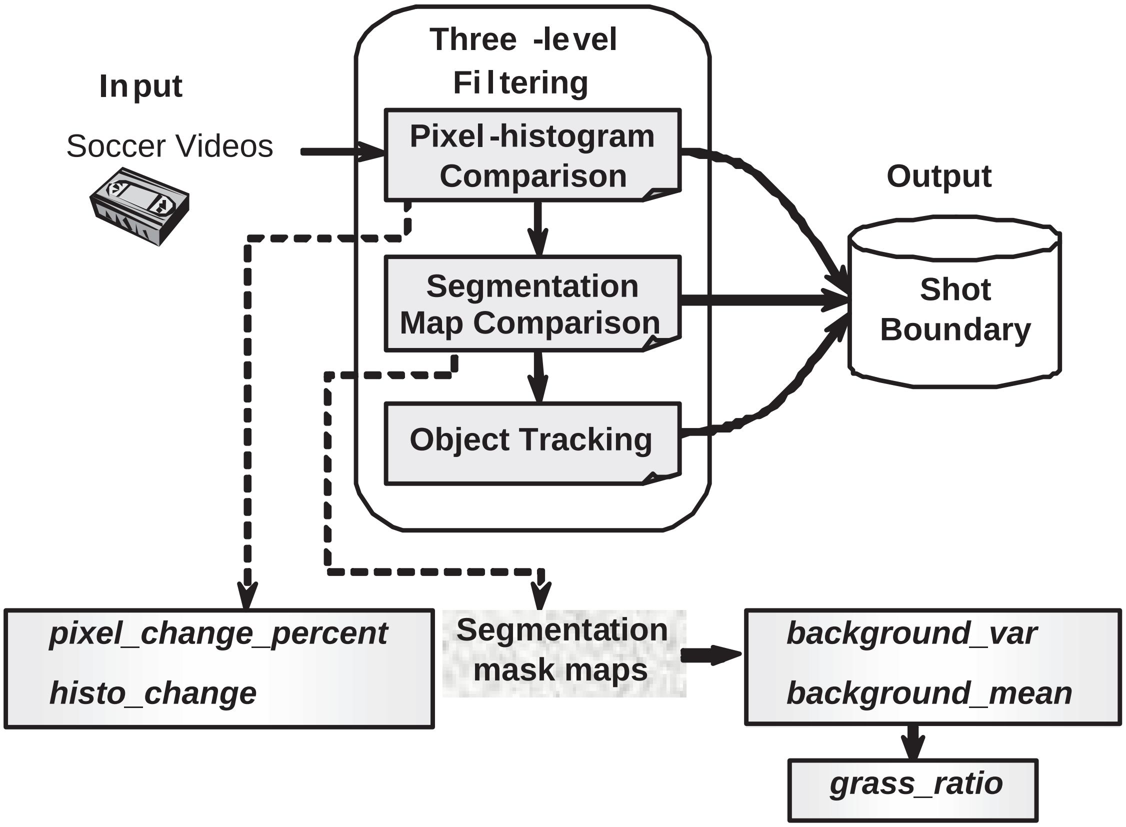


Figure 5.3: Framework architecture.

utive frames within a shot. Similarly, histo_change represents the mean value of the frame-to-frame histogram differences in a shot. Obviously, as illustrated in Fig. 5.3, pixel_change and histo_change can be obtained simultaneously and at a low cost during the video shot detection process. As mentioned earlier, both features are important indicators of camera motion and object motion.

In addition, as mentioned earlier, by using the WavSeg unsupervised object segmentation method, the significant objects or regions of interests as well as the segmentation mask map of a video frame can be automatically extracted. In such a way, the pixels in each frame are grouped into different classes (in this case, 2 classes called class1_region and class2_region marked with gray and white, respectively, as shown in Fig. 5.2(b)) for region-level analysis. Intuitively, features class1_region_mean (class2_region_mean) and class1_region_var (class2_region_var) represents the mean value and standard deviation of the pixels that belong to class1_region (class2_region) for the frames in a shot. In this study, the calculation of such features is conducted in the HSI (Hue-Saturation-Intensity) color space.

Audio Feature Descriptor Extraction

Extracting effective audio features is essential in achieving a high distinguishing power in audio content analysis for video data. A variety of audio features have been proposed in the literature for audio track characterization [72][125]. Generally, they fall into two categories: time domain and frequency domain. Considering the requirements of specific applications, the audio features may be extracted at different granularities such as frame-level and clip-level. In this section, several features are described that are especially useful for classifying audio data.

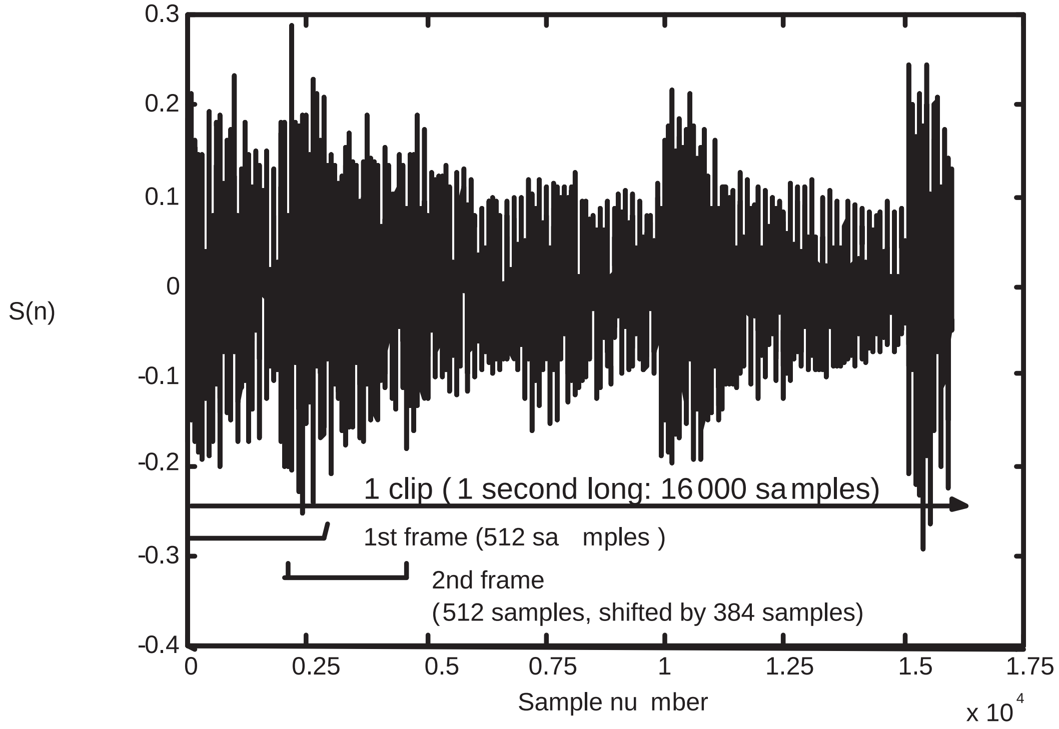
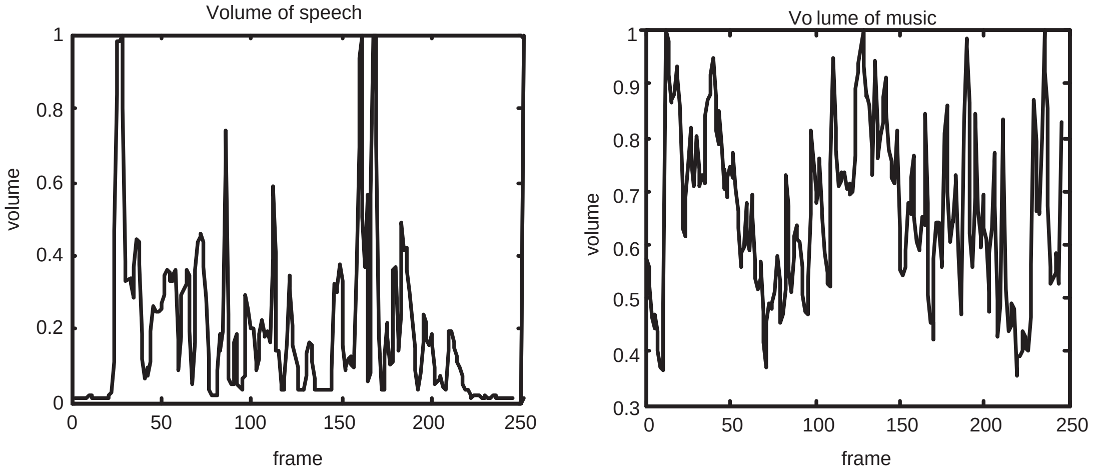


Figure 5.4: Clip and frames used in feature analysis.



(a) Speech

(b) Music

Figure 5.5: Volume of audio data.

The proposed framework exploits both time-domain and frequency-domain audio features. In order to investigate the comprehensive meaning of an audio track, the features representing the characteristics of a comparable longer period are necessary. In this work, both clip-level features and shot-level features are explored, which are obtained via the analysis of the finer granularity features such as frame-level features. In this framework, the audio signal is sampled at 16,000 Hz, i.e., 16,000 audio samples are generated for a one-second audio track. The sample rate is the number of samples of a signal that are taken per second to represent the signal digitally. An audio track is then divided into clips with a fixed length of one second. Each audio feature is first calculated on the frame-level. An audio frame is defined as a set of neighboring samples which lasts about 10∼40ms. Each frame contains 512 samples shifted by 384 samples from the previous frame as shown in Fig. 5.4. A clip thus includes around 41 frames. The audio feature analysis is then conducted on each clip (e.g., an audio feature vector is calculated for each clip).

The generic audio features utilized in this framework can be broadly divided into three groups: volume related, energy related, and Spectrum Flux related features.

- Feature 1: Volume

  Volume is one the most frequently used and the simplest audio features. As an indication of the loudness of sound, volume is very useful for soccer video analysis. Volume values are calculated for each audio frame. Fig. 5.5 depicts samples of two types of sound tracks: speech and music. For speech, there are local minima which are close to zero interspersed between high values. This is because when we speak, there are very short pauses in our voice. Consequently, the normalized average volume of speech is usually lower than that of music. Thus, the volume feature will help not only identify exciting points in the game, but also distinguish commercial shots from regular soccer video shots. According to these observations, four useful clip-level features related to volume can be extracted: 1) average volume (volume_mean), 2) volume_std, the standard deviation of the volume, normalized by the maximum volume, 3) volume_stdd, the standard deviation of the frame to frame difference of the volume, and 4) volume_range, the dynamic range of the volume, defined as $(max(v) - min(v))/max(v)$.

- Feature 2: Energy

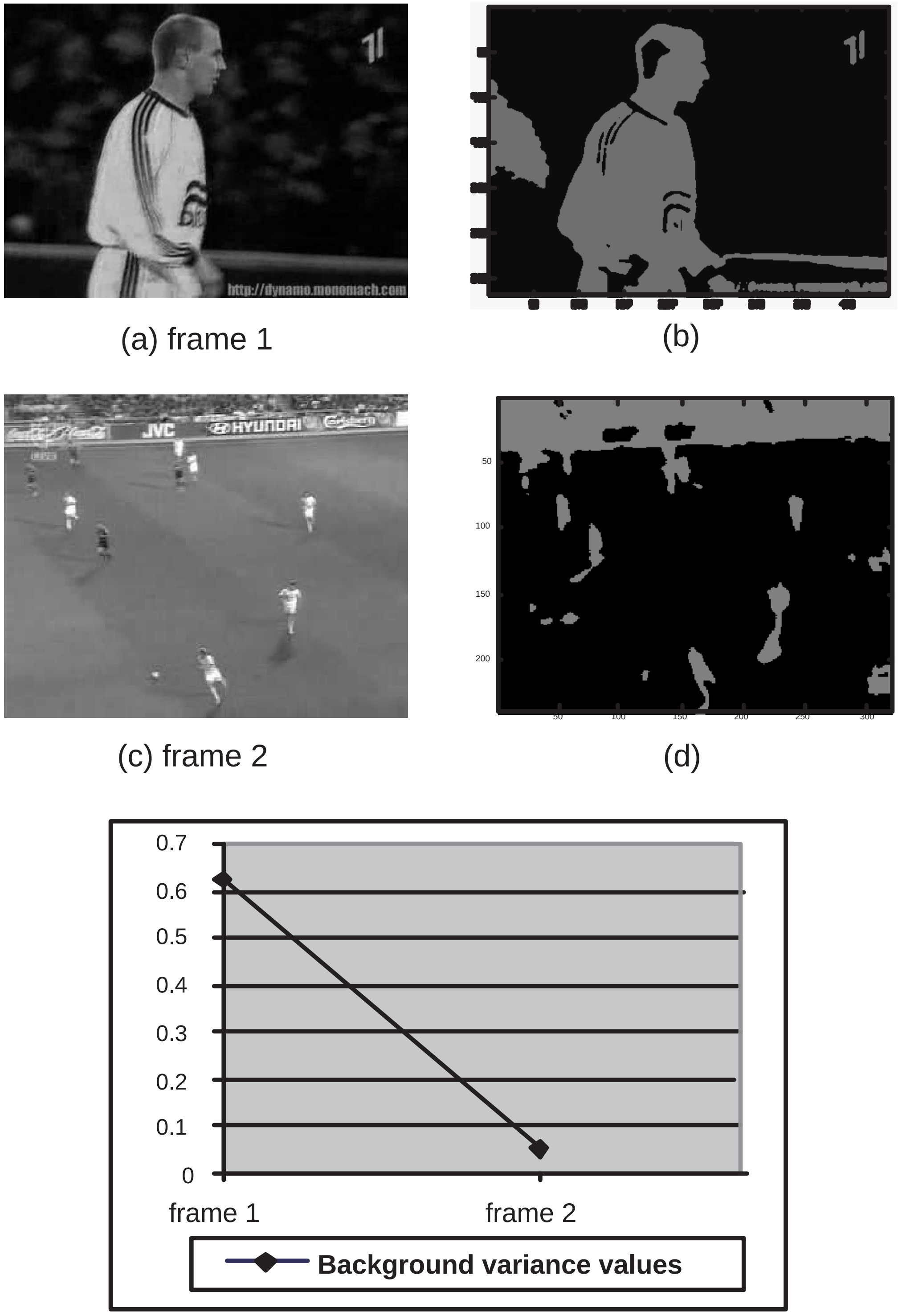
  Short time energy means the average waveform amplitude defined over a specific time window. In general, the energy of an audio clip with music content has a lower dynamic range than that of a speech clip. The energy of a speech clip changes frequently from high peaks to low peaks. Since the energy distribution in different frequency bands varies quite significantly, energy characteristics of sub-bands are explored as well. Four energy sub-bands are identified, which cover, respectively, the frequency interval of 1Hz-(fs/16)Hz, (fs/16)Hz-(fs/8)Hz, (fs/8)Hz-(fs/4)Hz and (fs/4)Hz-(fs/2)Hz, where fs is the sample rate. Compared to other sub-bands, sub-band1 (1Hz-(fs/16)Hz) and sub-band3 ((fs/8)Hz-(fs/4)Hz) appear to be most informative. Several clip-level features over sub-band1 and sub-band3 are extracted as well. Thus, the following energy-related features are extracted from the audio data: 1) energy_mean, the average RMS (Root Mean Square) energy, 2) The average RMS energy of the first and the third subbands, namely sub1_mean and sub3_mean, respectively, 3) energy_lowrate, the percentage of samples with the RMS power less than 0.5 times of the mean RMS power, 4) The energy-lowrates of the first sub-band and the third band, namely sub1_lowrate and sub3_lowrate, respectively, and 5) sub1_std, the standard deviation of the mean RMS power of the first sub-band energy.

- Feature 3: Spectrum Flux

  Spectral Flux is defined as the squared difference of two successive spectral amplitude vectors. Spectrum flux is often used in quick classification of speech and non-speech audio segments. In this study, the following Spectrum Flux related features are explored: 1) sf_mean, the mean value of the Spectrum Flux, 2) the clip-level features sf_std, the standard deviation of the Spectrum Flux, normalized by the maximum Spectrum Flux, 3) sf_stdd, the standard deviation of the difference of the Spectrum Flux, which is also normalized, and 4) sf_range, the dynamic range of the Spectrum Flux. Please note that the audio features are captured at different granularities: frame-level, clip-level, and shot-level, to explore the semantic meanings of the audio track. Totally, 15 generic audio features are used (4 volume features, 7 energy features, and 4 Spectrum Flux features) to form the audio feature vector for a video shot.

5.2.2 Mid-level Data Representation

Low-level audio-visual feature descriptors can be acquired directly from the input video data in (un)compressed domain. However, due to their limited capabilities in presenting the semantic contents of the video data, it is a traditionally open problem to establish the mappings between the low-level feature descriptors and semantic events.



(e)

Figure 5.6: (a) a sample frame from a goal shot (global view); (b) a sample frame from a close-up shot; (c) object segmentation result for (a); (d) object segmentation result for (b); (e) background variance values for frame 1 and frame 2

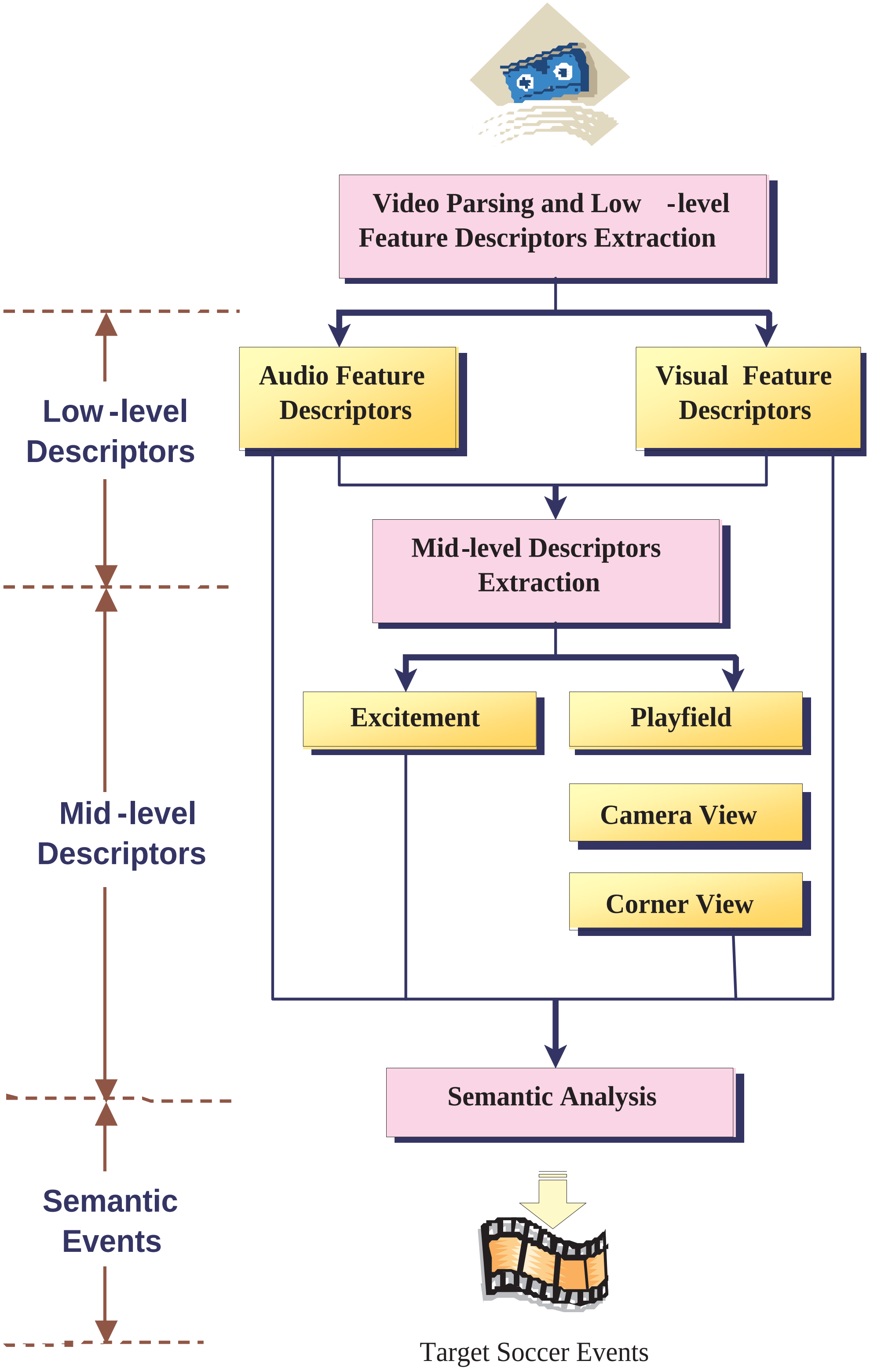
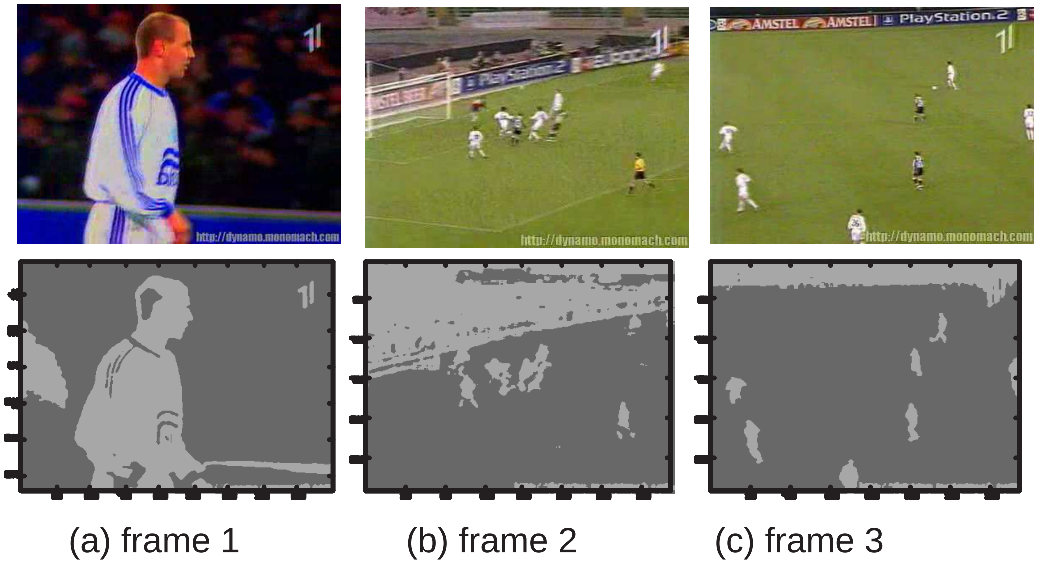

Video Parsing and Low -level Feature Descriptors Extraction
Audio Feature Descriptors
Visual Feature Descriptors
Low -level Descriptors
Mid-level Descriptors Extraction
Excitement
Playfield
Camera View
Corner View
Mid -level Descriptors
Semantic Analysis
Semantic Events
Target Soccer Events

Figure 5.7: Idea of Mid-level Data Representation.

Building mid-level descriptions is therefore considered as an effective attempt to address this problem [137]. Therefore, once the proper low-level visual and audio features have been extracted, a group of mid-level descriptors are introduced which are deduced from low-level feature descriptors and are motivated by high-level inference. Such mid-level descriptors offer a reasonable tradeoff between the computational requirements and the resulting semantics. In addition, the introduction of the mid-level descriptors allows the separation of sports-specific knowledge and rules from the extraction of low-level feature descriptors and offers robust and reusable representations for high-level semantic analysis using customized solutions. The aforementioned idea is illustrated in Fig. 5.7. In this work, four kinds of mid-level descriptors are extracted to represent the soccer video contents.

Field Descriptor

In sports video analysis, playfield detection generally serves as an essential step in determining other critical mid-level descriptors as well as some sport highlights. In soccer video analysis, the issue is defined as grass area detection. As can be seen from Fig. 5.6 (a)-(b), a large amount of grass areas are present in global shots (including goal shots), while fewer or hardly any grass areas are present in the mid- or the close-up shots (including the cheering shots following the goal shots). However, it is a challenge to distinguish the grass colors from others because the color values may change under different lighting conditions, different play fields, different shooting scales, etc. The method proposed in [38] relies on the assumption that the play field is always green in order to extract the grass areas, which is not always true for the reasons mentioned above. In [113], the authors addressed this issue by building a table with candidate grass color values. As a more robust solution, the work in [32] proposed to use the dominant color based method to detect grass areas, which does not assume any specific value for the play field color. However, the initial field color in [32] is obtained in the learning process



(a) frame 1 (b) frame 2 (c) frame 3

Figure 5.8: Three example video frames and their segmentation mask maps.

by observing only a few seconds of a soccer video. Thus, its effectiveness largely depends on the assumption that the first few seconds of video are mainly field play scenes. It also assumes that there is only a single dominant color indicating the play field, which fails to accommodate variations in grass colors caused by different camera shooting scales and lighting conditions. In this study, an advanced strategy in grass area detection is adopted, which is conducted in three steps as given below.

- Step 1: Extract possible grass areas

  The first step is to distinguish the possible grass areas from the player/audience areas, which is achieved by examining the segmentation mask maps of a set of video frames, S, extracted at 50-frame interval for each shot. Compared to the non-grass areas, the grass areas tend to be much smoother in terms of color and texture distributions. Motivated by this observation, for each frame, a comparison is conducted between class1_region_var and class2_region_var, where the class with the smaller value is considered as the background class and its mean value and standard deviation are thus called background_mean and background_var, respectively. Correspondingly, the other class is regarded as foreground. Three sample video frames and their corresponding segmented mask maps are shown in Fig. 5.8, where the background and foreground areas are marked with dark gray and light gray, respectively.

As shown in Fig. 5.8, the grass areas tend to correspond to the background areas (see Figs. 5.8(b) and 5.8(c)) due to the low variance values. On the other hand, for those frames with no grass area (e.g., Fig. 5.8(a)), the background areas are much more complex and may contain crowd, sign board, etc., which results in higher background_var values. It is worth mentioning that all the features used in this work are normalized in the range of [0,1]. Therefore, a background area is considered as a possible grass area if its background_var is less than T_b, which can be determined by statistical analysis of the average variation of field pixels. Thus, grass_ratio_approx is defined as the ratio of the possible grass area over the frame size. Note that the value of grass_ratio_approx is an approximate value, which will be utilized in step 2 to select the reference frames and will be refined in step 3.

- Step 2: Select reference frames to learn the field colors

  The reference frames are critical in learning the field colors. An ideal set of reference frames should contain a relatively high percentage of play field scenes with large grass ratios. Therefore, instead of selecting the reference frames blindly, in this work, the reference frames are selected from the shots with their grass_ratio_approx greater than T_{grass}. Here T_{grass} is set to the mean value of the grass_ratio_approx across the whole video clip. Since the feature background_mean represents the mean color value of each possible grass area, the color histogram is then calculated over the pool of the possible field colors collected for a single video clip. The actual play field colors are identified around the histogram peaks. It is not sufficient to have a single dominant color corresponding to the field color for a soccer video.

Hence, multiple histogram peak values are used as the field colors to accommodate the varied field/lighting conditions and the color differences caused by different shooting scales using the approach discussed in [11].

- Step 3: Refine grass_ratio values

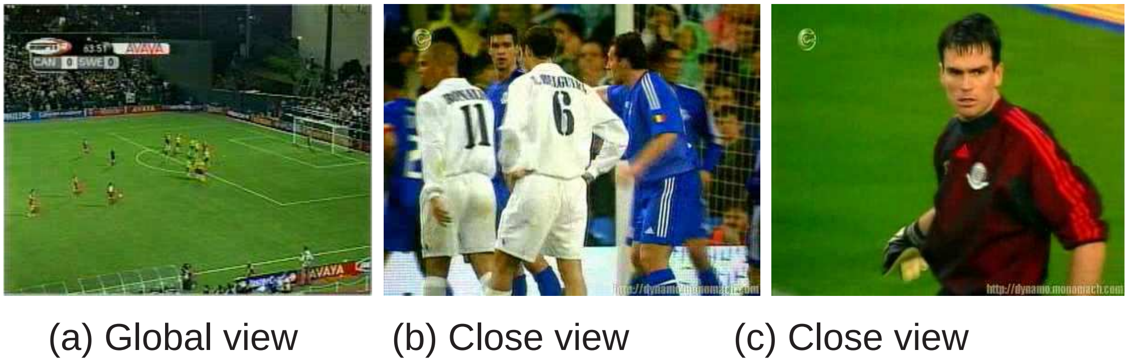
  Once the play field colors are identified, the refining of the grass_ratio value for a video shot is straightforward. In brief, for each segmented frame in S, the field pixels are detected from the background areas and thus its grass_ratio_approx can be refined to yield the accurate shot-level grass_ratio values. Note that since the background areas have been detected in step 1, the computational cost of this step is quite low. Similarly, data normalization is done within each video sequence.

In summary, the detected grass_ratio acts as the field descriptor which facilitates the extraction of some other mid-level descriptors (i.e., camera view descriptor and corner view descriptor to be discussed below) as well as the semantic event detection. It is also worth noting that by deducing grass_ratio at the region-level, the problem is resolved that the non-grass areas (e.g., sign boards, player clothes, etc.) may have close-to grass color, which fails to be addressed in most of the existing works. It is worth noting that the proposed grass area detection method is unsupervised and the grass values are learned through unsupervised learning within each video sequence. Therefore it is invariant to different videos.

The major theoretical advantages of the proposed approach are summarized as follows.

- The proposed method allows the existence of multiple dominant colors, which is flexible enough to accommodate variations in grass colors caused by different camera shooting scales and lightning conditions.

- In the learning process, the proposed method adopts an automated and robust approach to choose the appropriate reference frames for the learning process.



(a) Global view (b) Close view (c) Close view

Figure 5.9: Example camera view.

Table 5.1: Camera view descriptor.

CVD	**Condition**	**Thresholds**
Outfield	$grass\_ratio < T_o$	$T_o = 0.05$
Global	$grass\_ratio \geq T_{g1} \wedge Max\_o < ST_{g2}$	$T_{g1} = 0.4$, $T_{g2} = 0.05$
Close	(grass_ratio $< T_{c1} \vee$ Max_o $> T_{c2}$) $\wedge$ grass_ratio $> T_o$	$T_{c1} = 0.4$, $T_{c2} = 0.25$, $T_0 = 0.05$
Medium	Otherwise	

Camera View Descriptor

In the literature, various approaches have been proposed for camera view classification. Most of the existing studies utilize grass ratio as an indicator of the view types, assuming that a global view (e.g., Fig. 5.9(a)) has a much greater grass ratio value than that of a close view (e.g., Fig. 5.9(b)) [120]. However, close view shots such as the one shown in Fig. 5.9(c) could have large grass ratio values. Thus, the use of grass ratio alone can lead to misclassifications. In contrast, Tong et al. [119] proposed to determine the shot view via the estimation of the object size in the view. However, it is usually difficult to achieve accurate object segmentation, especially with the existence of object occlusions as shown in Fig. 5.9(b).

To address these issues, in this study, a hierarchical shot view classification scheme is proposed as illustrated in Fig. 5.10. As can be seen from this figure, grass ratio values act as the major criterion in differentiating the outfield views and infield views. Then

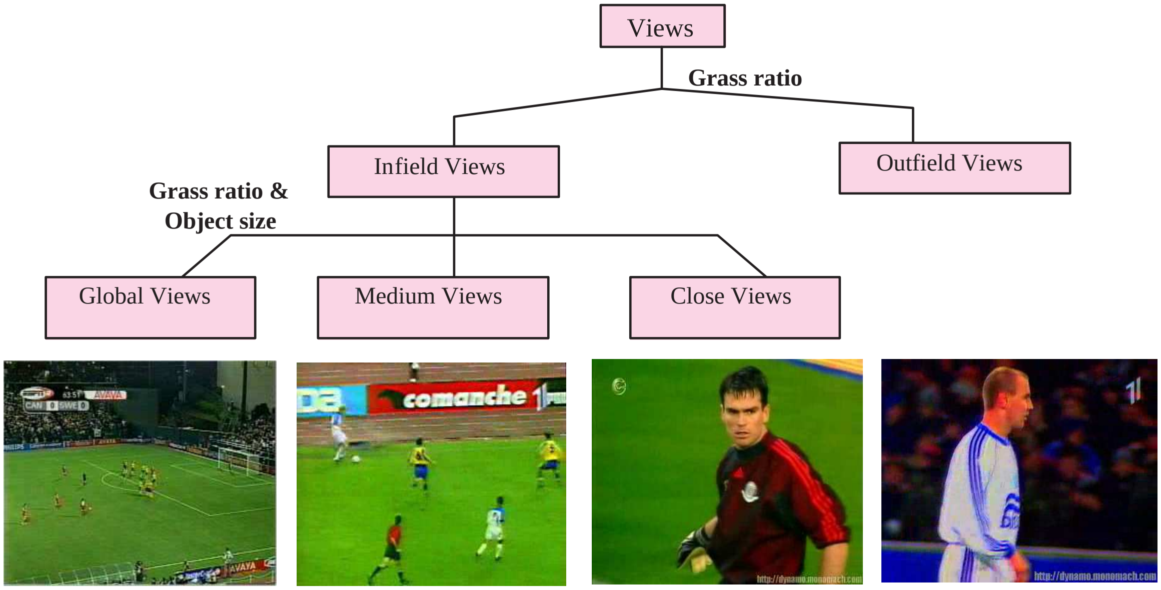


Figure 5.10: Hierarchical shot view.

the infield views are further categorized into close views, medium views and global views using the grass ratio value coupled with the object size in the playfield. The reasons for such a setting are twofold. First, a further classification of outfield views normally fails to yield more useful information to serve users' interests. Thus, to simplify the problem, only the infield views are further analyzed. Second, it is relatively easy to detect the grass area as opposed to the object detection due to its homogeneous characteristic, and the proposed playfield segmentation scheme can yield quite promising results. Therefore, the grass ratio value serves as the primary differentiating factor with the facilitation of roughly estimated foreground object size in the playfield area. In brief, the foreground object with the maximal size in the field is identified, and Max_O is calculated to denote the ratio of its area versus the frame size. The camera view descriptor is then defined as shown in Table 5.1.

Currently, the thresholds are defined empirically. A statistical analysis or data classification approach might help in this manner.

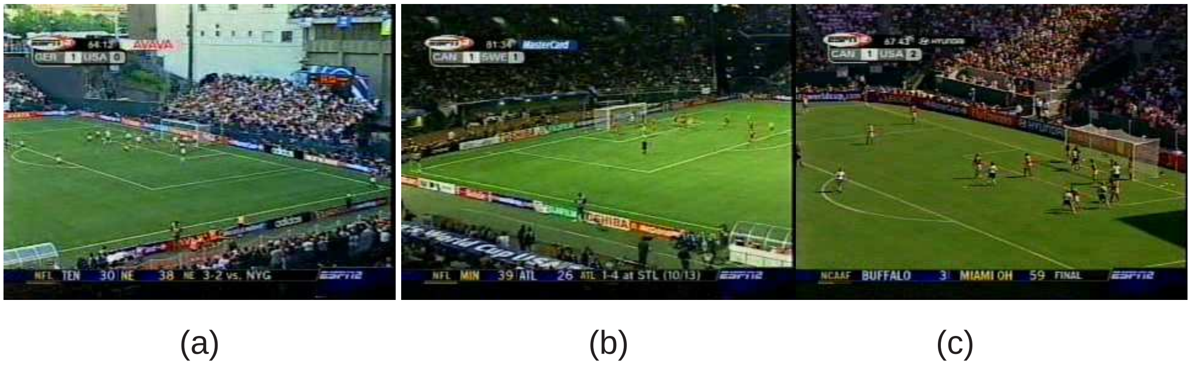


Figure 5.11: Example corner events.

Corner View Descriptor

A corner view is defined as to have at least one corner visible in the scene. The reason for defining the corner view lies in the fact that a large number of exciting events belong to corner events such as corner-kicks, free-kicks near the penalty box, and line throws from the corner (see examples in Fig. 5.11), which grants the opportunity for one team to dominate the other and possibly leads to a goal event. It is obvious that the identification of corner views can greatly benefit corner event detection. In [137], the so-called shape features ls (slope of top left boundary), rs (slope of right boundary), bs (slope of bottom boundary) and cp (corner position) were defined. However, it is not discussed in the paper as how such features can be extracted. In fact, due to the complicated visual contents in the videos, it remains an open issue to detect the field lines accurately, not to mention the attempt to label the field line correctly as left, right or bottom boundary. In this study, a simpler yet effective approach for corner views detection proposed in our previous work [10] is adopted. The basic idea is that though the minor discrepancy or noise contained in the segmentation mask map might deteriorate the performance of the direct identification of the corner point, the adverse effect of the bias can be compensated and thus reduced by intelligently examining the size of the grass area and audience area for the purpose of corner point detection. Detailed discussion can be found in [10].

Excitement descriptor

Different from the visual effects, the sound track of a video does not necessarily show any significant change at the shot boundary. To avoid the loss of actual semantic meanings of the audio track, the audio mid-level representation called excitement descriptor is defined to capture the excitement of the crowd and commentator in sport videos. Such an excitement is normally accompanied with or is the result of certain important events. The excitement descriptor is captured in a three-stage process. First, the audio volume feature is extracted at the clip-level. Here, an audio clip is defined with a fixed length of one second, which usually contains a continuous sequence of audio frames. Secondly, a clip with its volume greater than the mean volume of the entire video is extracted as an exciting clip. Finally, considering that such excitement normally lasts a period of time as opposed to other sparse happenings of high-volume sound (such as environmental sound) or noises, a time period with multiple exciting clips is considered to define the excitement descriptor. Here the time period is of fixed length and can be determined by adopting our previously proposed temporal pattern analysis algorithm [24]. In this study, for each shot, a time period of 6 sec is examined which includes the last 3-clip portion of this shot (for short, last_por) as well as the first 3-clip portion of its consecutive shot (for short, nextfirst_por). If one or more exciting clip(s) is detected in each of these 3-sec portions, vol_last (vol_nextfirst) is defined to record the maximum volume of last_por (nextfirst_por) and the excitement descriptor is the summation of vol_last and vol_nextfirst.

5.3 Video Indexing and Retrieval

Indexing video data is essential for providing content-based retrieval, which tags video clips when the system inserts them into the database. As discussed earlier, one focus of this chapter is event-level indexing. Therefore, with the proper video data representation, the next step is to effectively infer the semantic events via integrating the multi-level data representation intelligently. In the literature, there are many approaches proposed using semantic rules defined based on domain knowledge. In [60], an event detection grammar was built to detect the "Corner Kick" and "Goal" soccer events based on the detection rules. However, these rules need to be completely studied and pre-defined for each target event prior to generating the grammar trees that are used to detect the events. For example, there were totally 16 semantic rules defined for the corner kick events in [60], which were derived by carefully studying the co-occurrence and temporal relationships of the sub-events (represented by semantic video segments) in soccer videos. However, there are several disadvantages to this approach: (1) The derived rules are based on limited observation of a small set of soccer videos (4 FIFA2002 videos), which may not hold true when applied to other soccer videos produced by different broadcasters. For example, the "PR" in [60] refers to the sub-event that one player runs to the corner just before the corner kick events. However, it is not a necessary pre-condition for corner kick events. (2) The classification performance of such rules largely depends upon the detection of sub-events. However, the detection of such sub-events is of the same difficulty level as, or sometimes even more difficult than, the target event. (3) The derivation of such a large set of rules requires considerable manual effort, which limits its generality.

In this section, a high-level semantic analysis scheme is presented to evaluate the effectiveness of using the multimodal multi-level descriptors in event detection. Generally speaking, the semantic analysis process can be viewed as a function approximation problem, where the task is to learn a target function f that maps a set of feature descriptors x (in this case, low-level and mid-level descriptors) to one of the pre-defined event labels y. The target function is called a classification model. Various data classification techniques, such as SVM, neural network, can be adopted for this purpose.

In this study, the decision tree logic is used for data classification as it possesses the capability of handling both numerical and nominal attributes. In addition, it is able to select the representative descriptors automatically and is mathematically less complex. A decision tree is a flow-chart-like tree structure that is constructed by recursively partitioning the training set with respect to certain criteria until all the instances in a partition have the same class label, or no more attributes can be used for further partitioning. An internal node denotes a test on one or more attributes (features) and the branches that fork from the node correspond to all possible outcomes of a test. Eventually, leaf nodes show the classes or class distributions that indicate the majority class within the final partition. The classification phase works like traversing a path in the tree. Starting from the root, the instances value of a certain attribute decides which branch to go at each internal node. Whenever a leaf node is reached, its associated class label is assigned to the instance. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner [45]. The information gain measure is used to select the test attribute at each node in the tree. The attribute with the highest information gain, which means that it minimizes the information needed to classify the samples in the resulting partitions and reflects the least 'impurity' in these partitions, is chosen as the test attribute for the current node. Numeric attributes are accommodated by a two-way split, which means one single breakpoint is located and serves as a threshold to separate the instances into two groups. The voting of the best breakpoint is based on the information gain value. More detailed discussions can be found in [45]. This framework adopts the C4.5 decision tree classifier [91].

As there is a wide range of events in the soccer videos, it is difficult to present extensive event detection results for all the event types. Therefore, in this study, two classes of events, goal events and corner events, are selected for performance evaluation since they significantly differ from each other in various aspects such as event pattern, occurrence frequency, etc. Before the decision tree based classification process starts, a feature set needs to be constructed, which is composed of a group of low-level and mid-level descriptors. In terms of the low-level feature descriptors, four visual descriptors (pixel_change, histo_change, background_mean, background_var) and 14 audio descriptors are used. In addition, four mid-level descriptors are included as well. Since most events are the result of past activities and might cause effects in the future, to capture the temporal characteristics, these mid-level descriptors are extracted for both the current shot and its two adjacent shots.

5.4 Experiments

The proposed framework has been rigorously tested on a large data set with over 7 hours (432 minutes) of soccer videos, which were collected from a variety of sources, such as European Cup 1998, World Cup 2002, and FIFA 2003, and are with different production/post-production styles, resolutions, and frame rates. The data set contains 3,043 video shots parsed by the aforementioned shot detection algorithm, where the number of corner event shots and goal shots is 145 and 29, respectively.

5.4.1 Experimental Settings

In the experiment, 2/3rds of the whole data set (called training data set) was used to train the model which was then tested by the remaining 1/3rd data (called testing data set). In order to avoid the overfitting problem in training the decision tree classifier, the so-called 5-fold cross-validation scheme is adopted for performance evaluation. The whole data set was randomly divided five times to obtain five different groups of training and testing data sets. Therefore, five models were constructed, where each of them was tested by its corresponding testing data. Such a scheme allows better estimations of the framework's capability in applying the learned event models to other unseen data.

Table 5.2: Performance of corner event detection.

	Corner Event #	**Identified**	**Missed**	**Misidentified**	**Recall**	**Precision**
Test 1	40	38	2	6	95.0%	86.4%
Test 2	46	45	1	9	97.8%	83.3%
Test 3	50	49	1	9	98.0%	84.5%
Test 4	43	42	1	7	97.7%	85.7%
Test 5	44	43	1	7	97.7%	86.0%
				Average	97.2%	85.2%

5.4.2 Event Detection Performance

The performance of the corner event detection is illustrated in Table 5.2. As shown in this table, the 'Missed' column indicates the number of false negatives, which means that the corner events are misclassified as noncorner events; whereas the 'Misidentified' column indicates the number of false positives, i.e., the noncorner events are identified as corner events. Consequently, recall and precision are defined as follows:

Recall = Identified/(Identified+Missed),

Precision = Identified/(Identified+Misidentified)

As can be seen from this table, the performance is very promising, especially for the recall rate which reaches over 97% by average. In fact, in sports event detection, the metric recall is normally weighted higher than precision as it is preferred to have all the targeted events detected even at the cost of including a small number of irrelevant shots. Also, a further check of the experimental results finds that most of the misidentified shots (i.e., false positives) are goal kicks/attempts whose event patterns are quite similar to that of the corner events. In fact, such events are usually considered as exciting events as well. In future work, the current framework can be further extended for goal attempt detection.

The framework is also tested upon the goal event detection and the performance is summarized in Table 5.3. As can be seen from this table, the results in terms of recall and

Table 5.3: Performance of goal event detection.

	Goal Event #	**Identified**	**Missed**	**Misidentified**	**Recall**	**Precision**
Test 1	11	10	1	1	91.7%	91.7%
Test 2	10	10	0	2	100.0%	83.3%
Test 3	12	11	1	2	92.3%	85.7%
Test 4	10	9	1	1	90.0%	90.0%
Test 5	11	10	1	1	90.9%	90.9%
				Average	93.0%	88.3%

precision are also quite satisfactory. It should be pointed out that the goal events account for less than 1% of the total data set. The rareness of the target events usually poses additional difficulties in the process of event detection. Through the cross-validation and multiple event detection, the robustness and effectiveness of the proposed framework is fully demonstrated in event detection.

5.5 Conclusions

In this chapter, shot boundary detection, data representation extraction and video indexing are discussed. In particular, a multi-level multimodal representation framework is proposed for event detection in field-sports videos. Compared with previous work in the sports video domain, especially in soccer video analysis, the proposed framework is unique in its systematic way of generating, integrating, and utilizing the low-level and mid-level descriptors for event detection to bridge the semantic gap. The extraction of low-level descriptors starts as early as in the shot detection phase, thus saving time and achieving better performance. Four generic and semi-generic mid-level descriptors (field descriptor, camera view descriptor, corner view descriptor, and excitement descriptor) are constructed from low-level visual/audio features, via a robust mid-level descriptor extraction process. In the high-level analysis, an event model is inferred from both low-level descriptors and mid-level descriptors, by using a decision tree based classification model, while most of the existing work infer only the relationship between the mid-level descriptors and the events. A large test soccer video data set, which was obtained from multiple broadcast sources, was used for experiments. Compared to the ground truth, it has been shown that high event detection accuracy can be achieved. Under this framework, domain knowledge in sports video is stored in the robust multi-level multimodal descriptors, which would be reusable for other field-sports videos, thus making the event detection less ad hoc. It is also worth pointing out that the proposed framework does not utilize any broadcast video features such as score board and other meaningful graphics superimposed on the raw video data. Such a framework is essential in the sense that it not only facilitates the video database in low-level and mid-level indexing, but also supports the high-level indexing for efficient video summarization, browsing, and retrieval.

It is worth noting that in order to bridge the semantic gap, in this study, the mid-level representation is captured with the assistance of some *a priori* or domain knowledge. In the next chapter, a set of automatic analysis techniques are proposed as an attempt to largely relax the dependence on the domain knowledge and human efforts by quantifying the contribution of temporal descriptors in field-sports video analysis.

CHAPTER 6

Automatic Knowledge Discovery for Semantic Event Detection

Generally speaking, the events detected by the existing methods (including the approaches introduced in Chapter 5) are semantically meaningful and usually significant to the users. The major disadvantage, however, is that most of these methods rely heavily on specific artifacts (so-called domain knowledge or *a priori* information) such as editing patterns in broadcast programs, which are generally explored, represented and applied with a great deal of costly human interaction. For instance, in [67], a set of thresholds need to be manually determined in order to associate the video sequences to the so-called visual descriptors such as "Lack of motion," "Fast pan," and "Fast zoom" whose temporal evolutions are in turn used for soccer goal detection. However, since the selection of timing for motion evolution is not scalable, its extensibility is highly limited. In our earlier studies [11], to cope with the challenges posed by rare event detection, a set of visual/audio clues and their corresponding thresholds were pre-defined based on the domain knowledge in order to prepare a "cleaned" data set for the data mining process. For the heuristic methods [120][147], the situation becomes even worse with the necessity of using a group of predefined templates or domain-specific rules. Such manual effort adversely affects the extensibility and robustness of these methods in detecting the different events in various domains.

With the ultimate goal of developing an extensible event detection framework that can be robustly transferred to a variety of applications, a critical factor is to relax the need for the domain knowledge, and hence to reduce the manual effort in selecting the representative patterns and defining the corresponding thresholds. Though such patterns usually play a critical role in video event detection, it is important to introduce an automatic process in developing an extensible framework, in terms of event pattern discovery, representation and usage. In response to this requirement, in this chapter, a novel temporal segment analysis method will be first discussed for defining the characteristic temporal segment and its associated temporal features with respect to the event unit (shot) [8]. Then in Section 6.2, a temporal association mining framework is proposed to systematically capture the temporal pattern from the temporal segment and automatically develop the rules for representing such patterns.

6.1 Temporal Segment Analysis for Semantic Event Detection

This section introduces a novel temporal segment analysis for semantic event detection. More specifically, the video data is considered as a time series $X = \{x_t, t = 1, ..., N\}$, where t is the time index and N is the total number of observations. Let $x_i \in X$ be an interesting event; the problem of event detection in this framework is decomposed into the following three subtasks. First, an event x_i should possess its own characteristics or feature set which needs to be extracted. In this framework, the audio-visual multimodal approach introduced in Chapter 5 is adopted since textual modality is not always available and is language dependent. Second, from the temporal evolution point of view, usually an event x_i is the result of past activities and might cause effects in the future as well. Therefore, an effective approach is required to explore the time-ordered structures (or temporal patterns) in the time series that are significant for characterizing the events of interest. It is worth noting that the ultimate purpose of this subtask is to explore and represent the temporal patterns automatically and to feed such valuable information intelligently to the next component. Finally, with the incorporation of both the feature set and the temporal patterns (or temporal feature set), advanced classification techniques are carried out to automatically detect the interesting events. It is in this step that the discovered temporal patterns are fully utilized for the purpose of event detection. The overview of the framework is illustrated in Fig. 6.1. As will be detailed, intelligent temporal segment analysis and decision tree based data mining techniques are adopted in this study to successfully fulfill these tasks with little human interference. Since the

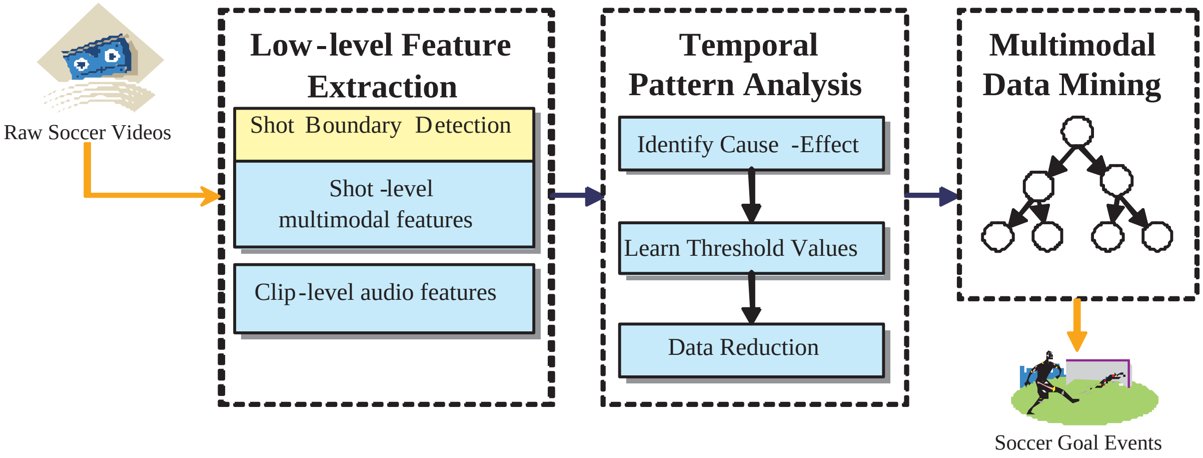


Figure 6.1: Overview of temporal segment analysis.

structure pattern of soccer videos is relatively loose and it is difficult to reveal high-level play transition relations by simply clustering the shots according to the field of view [67], this chapter will focus on the application of soccer goal event detection on a large collection of soccer videos to demonstrate the effectiveness of the proposed framework.

6.1.1 Temporal Pattern Analysis

As discussed in [90], given a training time series $X = \{x_t, t = 1, ..., N\}$, the task of defining a temporal pattern p is to identify $X_t = \{x_{t-(Q-1)\tau}, ..., x_{t-\tau}, x_t\}$ from X, where x_t represents the present observation and $x_{t-(Q-1)\tau}, ..., x_{t-\tau}$ are the past activities. With the purpose of predicting the future events in [90], their goal is to capture the temporal patterns which occur in the past and are completed in the present, with the capability of forecasting some event occurring in the future. However, because of noise, the temporal pattern p does not perfectly match the time series observation in X. The rule (i.e., temporal pattern) intended to discover might not be obtained directly by using the actual data points (i.e., the observations) in X. Instead, a temporal pattern cluster is required to capture the variability of a temporal pattern. Here, a temporal pattern cluster P is defined as a neighborhood of p, which consists of all the temporal observations within a certain distance d of p with respect to both its value and the time of occurrence.

This definition provides the theoretic basis for this proposed temporal pattern analysis algorithm. However, many of the existing temporal analysis approaches [34][90] focused on finding the significant pattern to predict the events. In contrast, as far as video event detection is concerned, the target is to identify the temporal patterns which characterize the events. Not only are the causes which lead to the event considered, but also the effects the event might have. Consequently, the problem can be formalized to identify a set of temporal observations

$$X_t = Y_{i=a}^{b}(x_{t+i\tau}), \tag{6.1}$$

which belong to the temporal pattern cluster P for a certain event occurring at x_t. Here, τ, called *granularity indicator*, represents the granularity level (i.e., shot-level, frame-level or clip-level in video analysis) at which the observations are measured, and the parameters a and b define the starting and ending positions of the temporal observations, respectively. Note that a and b are not limited to positive integers in the sense that there might not be any cause for a certain event at x_t or the event might be ended with no effect on the later observations. Furthermore, X_t might not contain x_t, if x_t contributes nothing in terms of the temporal evolution for the event occurring at x_t.

Moreover, the approaches proposed in [34][90] are targeted to a direct decision of the “eventness” of the testing units. A genetic algorithm or fuzzy objective function therefore needs to be applied to search for the optimal heterogeneous clusters in the augmented phase space. In contrast, the purpose of the temporal pattern analysis, which lies in two aspects as discussed earlier, differs substantially. Consequently, a new methodology is adopted in this framework for two reasons. First, with a high-dimensional feature set as multimedia data usually have, the genetic algorithm is extremely time consuming and thus becomes infeasible, especially for sports video whose relevance drops significantly after a relatively short period of time. Second, the algorithm in [34] requires that the observations are extracted at the same granularity level τ, whereas for video analysis,

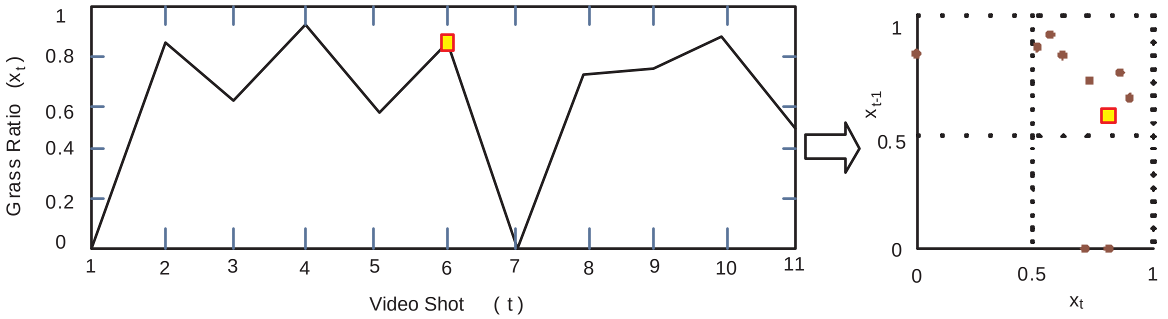


Figure 6.2: An example two-dimensional temporal space for a time series data.

the visual and audio features are normally examined at different levels.

In order to explore the most significant temporal patterns, three important concepts must be addressed as follows.

- Concept 1. How to define the similarity relationships (or distance function) among the time series observations.

- Concept 2. How to determine appropriate temporal segmentation, that is, how to determine the values of parameters a, b and τ in Eq. (6.1), which define the time window (its size and position) where the temporal pattern is presented.

- Concept 3. How to define the threshold d in order to determine the temporal pattern cluster.

To address these concepts, a temporal space is constructed in the proposed framework. Here, the temporal space is defined as a $(b-a)$ dimensional metric space into which X_t is mapped. The temporal observations in X_t can be mapped to a point in the dimensional space, with $x_{t+a\tau}, ..., x_{t+b\tau}$ being the coordinate values, whereas the "eventness" of unit x_t is assigned as the label of the temporal observations X_t. Fig. 6.2 gives an example. Assume the yellow square marker indicates a certain event of interest. The left side of the figure shows the grass ratios extracted from the consecutive video shots; whereas the right figure shows the corresponding two-dimensional temporal space for this time series

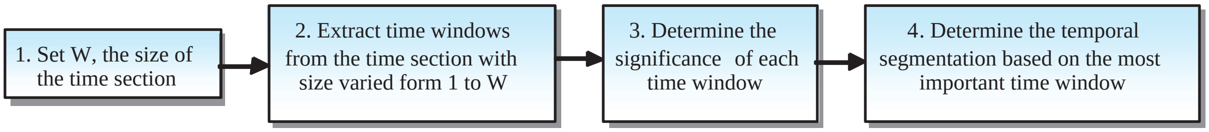

1. Set W, the size of the time section
2. Extract time windows from the time section with size varied form 1 to W
3. Determine the significance of each time window
4. Determine the temporal segmentation based on the most important time window

Figure 6.3: Overview of the algorithm for temporal segmentation.

data. The coordinates of each point, denoted as (x_t, x_{t-1}) in the right figure, represent the grass ratio values obtained at time t and time $t-1$, respectively, where the event label of the unit xt in the left figure is treated as the label of the point (x_t, x_{t-1}) in the right figure.

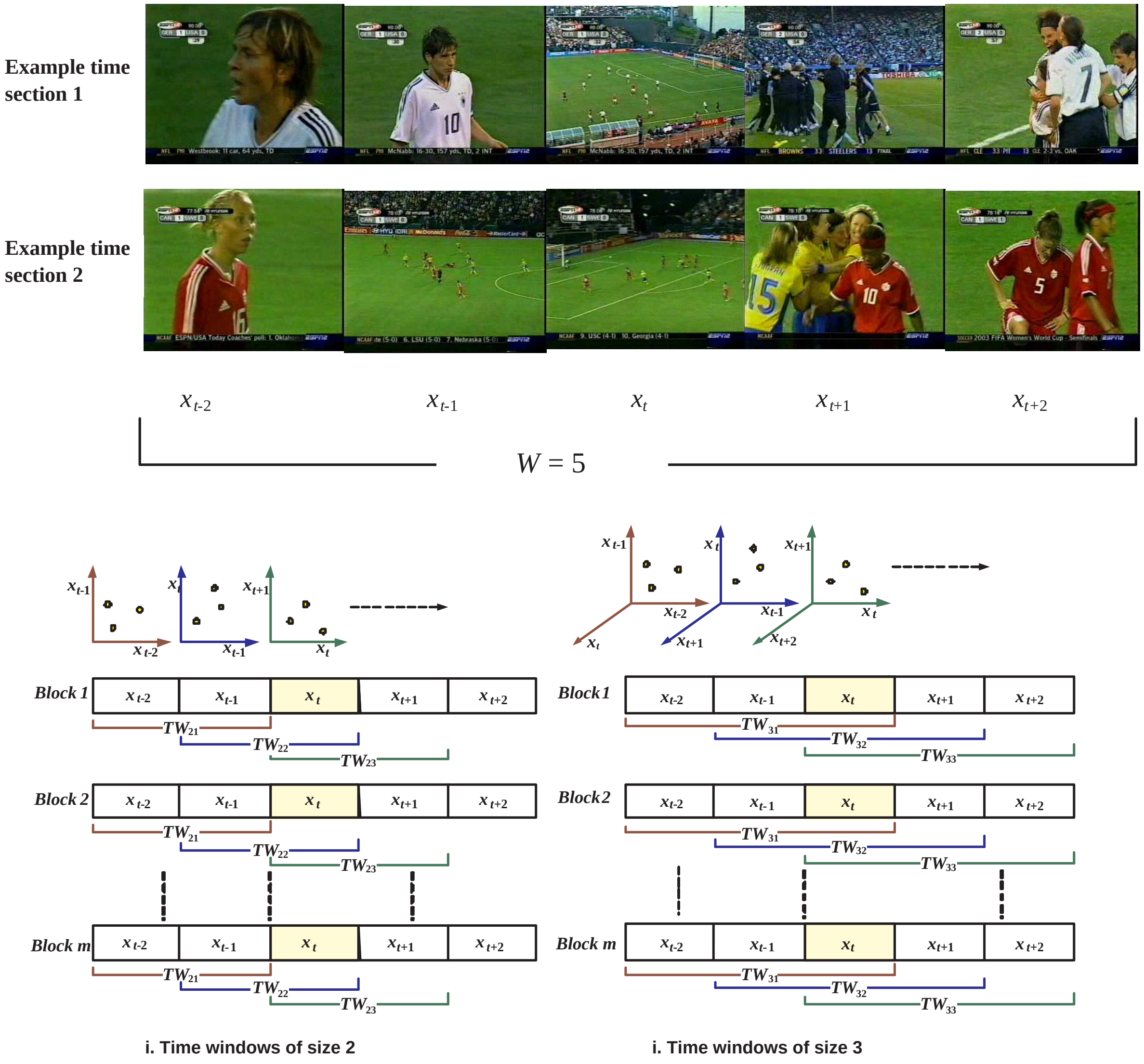
Concept 1. Distance Function

With the introduction of the temporal space, the problem raised in Concept 1 is converted to the calculation of the distances among the points in a certain space, where various distance functions (e.g., *Euclidean* or *Manhattan* distance functions) can be easily applied. In fact, as discussed in [90], it is generally considered an effective approach to adopt the space transition idea and the distance metrics in the case when the direct calculation of the distance is too complicated to perform in the original space.

Concept 2. Temporal Segmentation

To determine appropriate temporal segmentation as discussed earlier, the event and visual features are defined at the shot-level. Therefore, τ is set to the shot-level in the temporal series for the visual features, whereas τ is defined at the clip-level for the audio features (as the reasons mentioned earlier). To define a and b, a time-window algorithm is developed for the visual features, which can be easily derived for audio features. The basic idea is that a significant temporal pattern should be able to separate the event units as far away as possible from the nonevent units, and in the meantime group the event units themselves as closely to each other as possible.

An overview of the algorithm is illustrated in Fig. 6.3 and a detailed discussion is given below.



i. Time windows of size 2

i. Time windows of size 3

(a) Two samples: key frames of the shots inside an example time section
(b) Example time windows

Figure 6.4: Time window algorithm.

1. Given a training data set $\{E, N\}$, $E = \{e_i, i > 0\}$ represents the set of event units, $N = \{n_j, j > 0\}$ is the set of nonevent units homogeneously sampled from the source data, and normally $|N| >> |E|$. Let W be the upper-bound size of the

searching window or the time section centered at e_i or n_j where the important temporal segmentation is searched. For soccer goal event detection, W is set to 5. Without loss of generality, it is assumed that a goal event might be directly caused by two past shots and significantly affect two future shots, as shown in the two examples in Fig. 6.4(a). In fact W can be set to any reasonably large number in this algorithm, as will be shown in the later steps. However, the larger the value of W, the greater is the computational cost that will be involved.

2. Define a set of time windows with various sizes w from 1 to W, which slide through the time sections and produce a group of observations which are mapped to the corresponding temporal space with dimension w (for example, Fig. 6.4(b) shows time windows of sizes 2 and 3). Here, blocks 1 to m represent a set of time sections with $m = |N| + |E|$, whereas TW_{21} and TW_{22} denote the first time window and the second time window when $w = 2$. Note that given a dimension w, it will have $W-w+1$ temporal spaces (as defined earlier in this section). Consequently, totally $(W \times (W+1)/2)$ temporal spaces will be generated with w varied from 1 to W.

3. A parameter S, called significance indicator, is defined to represent the importance of each time window.

$$S = \sum_{i \in E} (\sum_{j \in N} D_{ij}) / \sum_{i \in E} (\sum_{k \in E, k \neq i} D_{ik}), \tag{6.2}$$

where D represents the distance between two points in the temporal space. S is defined as the ratio of the distance between every point in the event set (E) and every point in the nonevent set (N) versus the distance among the points in E. The change of the window size in step 2 results in the alteration of data dimensions. It is well-known that as the dimensionality of the data increases, the distances between the data points also increase [80]. Therefore, the capability of Eq. (6.2) is briefly discussed in handling the performance comparisons among the time windows with various sizes. Let δ_{ij} and δ_{ik} be the increments caused by the introduction of the new dimensions. After increasing the size of time window, we get

$$S' = \sum_{i \in E}[\sum_{j \in N}(D_{ij} + \delta_{ij})]/\sum_{i \in E}[\sum_{k \in E, k \neq i}(D_{ik} + \delta_{ik})] \tag{6.3}$$

$$S' = \frac{\sum_{i \in E}(\sum_{j \in N} D_{ij}) + \sum_{i \in E}(\sum_{j \in N} \delta_{ij})}{\sum_{i \in E}(\sum_{k \in E, k \neq i} D_{ik}) + \sum_{i \in E}(\sum_{k \in E, k \neq i} \delta_{ik})}$$

$$= \frac{\sum_{i \in E}(\sum_{j \in N} D_{ij})(1 + \sum_(i \in E)(\sum_{j \in N} \delta_{ij})/\sum_(i \in E)(\sum_{j \in N} D_{ij}))}{\sum_{i \in E}(\sum_{k \in E, k \neq i} D_{ik})(1 + \sum_{i \in E}(\sum_{k \in E, k \neq i} \delta_{ik})/\sum_{i \in E}(\sum_{k \in E, k \neq i} D_{ik}))} \tag{6.4}$$

In other words,

$$S' = S \times \frac{(1 + R_{EN})}{(1 + R_E)} \tag{6.5}$$

Here, R_{EN} represents the incremental rate of the distance between the units in E and N, and R_E is the incremental rate of the distance among the units in E, respectively. It can be observed from Eq. (6.5) that when $S' > S$, the new dimension possesses a greater impact on R_{EN} than on R_E (i.e., $R_{EN} > R_E$).

4. Also, as the significance indicator S increases, so does its importance as a time window. Therefore, a and b are determined by the time window(s) with the greatest S. If there is a tie, then it is broken by the preferences in the following order: 1) choosing a window with a smaller size, which will require less computational cost in the later processes; and 2) selecting a window closer to x_t as it is generally the case that the nearby activities have a relatively higher affinity with x_t.

Since the grass ratio is among the most important visual features for many sports video, the above-mentioned algorithm is carried out for the grass ratio feature in the proposed framework. It is worth mentioning, though, that without any *a priori* information,

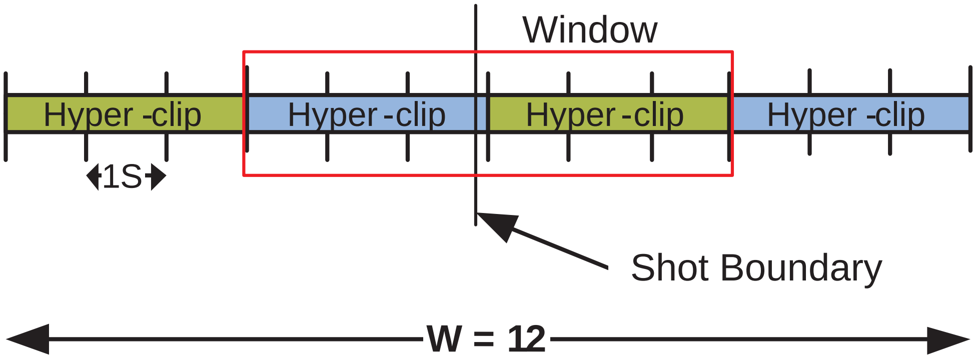


Figure 6.5: Time window for the Volume feature.

the same procedure can be carried out for other visual features as well and the one with the largest S value contains the most important temporal pattern.

As for the audio track, sound loudness is one of the simplest and most frequently used features in identifying the excitement of the crowd and commentator in sports videos. The time-window algorithm can be applied on sound loudness as well with minor revisions. Specifically, as τ is set to the clip-level, the size of the time section W is usually set to a larger value than the one used for shot-level visual features. In the current implementation, W is set to 12, as shown in Fig. 6.5. Here, the ending boundary of shot e_i or n_j is set as the center of the time section or as closely to the center as possible. In real implementations, the latter occurs more frequently since the shot boundary and clip boundary usually do not match each other. However, as can be seen from the time window algorithm, the computational cost increases by order $O(W^2)$. Therefore, for the sake of efficiency, the time-window algorithm can be revised to apply on the hyper-clip level. The 12-clip time section is broken into a set of hyper-clips. Here, a hyper-clip is defined as a unit that consists of three consecutive clips in this framework, and is represented by its statistical characteristics such as *volume_mean* (the mean volume of the hyper-clip) and *volume_max* (the max volume of the hyper-clip).

Formally, assume that C_h is a hyper-clip consisting of three consecutive clips c_1, c_2, and c_3 whose volume values are v_1, v_2, and v_3, respectively. We have

$$volume\_mean = mean(v_1, v_2, v_3). \tag{6.6}$$

$$volume\_max = max(v_1, v_2, v_3). \tag{6.7}$$

By applying the time-window algorithm, for each shot e_i or n_j, the most important time window (marked by the red rectangle in Fig. 6.5) in terms of the sound loudness can be obtained. This time window consists of two hyper-clips L_h and H_h (as shown in Fig. 6.5). The *volume_mean* and *volume_max* features in L_h are named as *last_mean* and *last_max*. Correspondingly, the *volume_mean* and *volume_max* features in H_h are called *nextfirst_mean* and *nextfirst_max*. Therefore, the significant temporal pattern for each shot can be represented by *last_mean*, *nextfirst_mean* and *volume_sum*. Here, *volume_sum* is defined as

$$volume\_sum = last\_max + nextfirst\_max, \tag{6.8}$$

which is introduced to magnify the pattern.

Concept 3. Temporal Pattern Cluster

With the purpose of data reduction, the third concept is more related to the problem of defining the threshold d to model the temporal pattern cluster. This is used to filter out inconsistent and noisy data and prepare a "cleaned" data set for the data mining process. The technique adopted is Support Vector Machines (SVMs). In a binary classification problem, given a set of training samples $\{(X_i, y_i), i = 1, 2, ..., n\}$, the i^{th} example $X_i \in R_m$ in an m-dimensional input space belongs to one of the two classes labeled by $y_i \in \{-1, 1\}$. The goal of the SVM approach is to define a hyperplane in a high-dimensional feature space Z, which divides the set of samples in the feature space such that all the points with the same label are on the same side of the hyperplane [116]. Recently, particular attention has been dedicated to SVMs for the problem of pattern recognition. As discussed in [80], SVMs have often been found to provide higher classification accuracies than other widely used pattern recognition techniques, such as the maximum likelihood and the multilayer perception neural network classifiers. Furthermore, SVMs also present strong classification capabilities when only few training samples are available.

However, in multimedia applications, data is represented by high dimensional feature vectors, which induces a high computational cost and reduces the classification speed in the context of SVMs [78]. Therefore, SVMs is adopted in the temporal pattern analysis step with the following two considerations. First, the classification is solely applied to a certain temporal pattern with few features. In the case of soccer goal event detection, SVMs is applied to the temporal patterns on the grass ratio feature only. Secondly, SVMs are capable of dealing with the challenges posed by the small number of interesting events. Currently, in this framework, SVM light [57] is implemented, which is an approach to reduce the memory and computational cost of SVMs by using the decomposition idea.

For soccer goal detection, it is identified that the most significant time window upon the grass ratio is TM_{33} by using the proposed time window algorithm, which means the grass ratios of the current shot and its two consecutive shots are important to characterize the goal event. The set of training examples fed into SVMs is defined as $\{(X_i, y_i), i = 1, 2, ..., n\}$, where the i^{th} example $X_i \in R^3$ belongs to one of the two classes labeled by $y_i \in \{-1, 1\}$ (i.e., nongoal or goal). Consequently, a SVM classifier can be learned to determine the threshold d automatically so as to classify the temporal pattern clusters of interest, which is thereafter applied upon the testing data for data reduction.

Table 6.1: Performance of goal event detection using temporal segment analysis.

	# of goals	**Identified**	**Missed**	**Misidentified**	**Recall**	**Precision**
Test 1	11	10	1	2	90.9%	83.3%
Test 2	14	12	2	4	85.7%	75.0%
Test 3	12	11	1	2	91.7%	84.6%
Test 4	11	11	0	2	100.0%	84.6%
Test 5	12	10	2	2	83.3%	83.3%
				Average	90.3%	82.2%

6.1.2 Experimental Results

The same experimental data set introduced in Chapter 5 was used in this section to testify the performance of the proposed temporal segment analysis approach. As discussed in section 6.1.1, the temporal segment analysis process by using the SVM light algorithm was carried out to produce a “cleaned” data set. An SVM classifier was trained by each of the training data sets in five groups and was applied upon the corresponding testing data set. After this step, the goal shots accounted for about 5% of the remaining data set, where many inconsistencies and irrelevant shots have been filtered out.

The resulting candidate pool was then passed to the decision tree based multimodal data mining process for further classification. Similarly, for each group, a decision tree model was built based on the “cleaned” training data set and was used to classify the corresponding testing data set in the candidate pool. The results are summarized in Table 6.1. The precision and recall values were computed for all the testing data sets in these five groups (denoted as Test 1, Test 2, etc.) to evaluate the performance of the proposed framework. Similarly, as defined in Chapter 5, the “Missed” column indicates a false negative, which means that the goal events are misclassified as nongoal events; whereas, the “Misidentified” column represents a false positive, i.e., the nongoal events that are identified as goal events. Consequently, precision and recall are defined as:

$Recall = Identified/(Identified + Missed)$,

$Precision = Identified/(Identified + Misidentified)$

From Table 6.1, it can been clearly seen that the results are quite encouraging in the sense that the average recall and precision values reach 90.3% and 82.2% respectively. To the best of our knowledge, this work is among the very few existing approaches in soccer video event detection whose performance is fully attested by a strict cross-validation method. In addition, compared to the work proposed in Chapter 5 which adopts mid-level representation with the assistance of the domain knowledge, the dependency on predefined domain knowledge in this framework is largely relaxed in the sense that an automatic process is adopted to discover, represent and apply the event specific patterns. Nevertheless, their performances are both very promising and quite close to each other, which demonstrates the effectiveness and robustness of this presented framework.

Conclusions

Event detection is of great importance for effective video indexing, summarization, browsing, and retrieval. However, due to the challenges posed by the so-called semantic gap issue and the rare event detection, most of the existing works rely heavily on domain knowledge with large human interference. To relax the need of domain knowledge, a novel framework is proposed for video event detection with its application to the detection of soccer goal events. Via the introduction of an advanced temporal segment analysis process, the representative temporal segment for a certain event can be explored, discovered and represented with little human effort. In addition, the multimodal data mining technique on the basis of the decision tree algorithm is adopted to select the representative features automatically and to deduce the mappings from low-level features to high-level concepts. As a result, the framework offers strong generality and extensibility by relaxing its dependency on domain knowledge. The experimental results over a large collection of soccer videos using the strict cross-validation scheme have demonstrated the effectiveness and robustness of the present framework.

6.2 Hierarchical Temporal Association Mining

As mentioned earlier, there are two critical issues in video event detection which have yet not been well studied.

1. First, normally a single analysis unit (e.g., shot) which is separated from its context has less capability of conveying semantics [148]. Temporal information in a video sequence plays an important role in conveying video content. Consequently, an issue arises as how to properly localize and model context which contains essential clues for identifying events. One of the major challenges is that for videos, especially those with loose content structure (e.g., sports videos), such characteristic context might occur at uneven inter-arrival times and display at different sequential orders. Some works have tried to adopt temporal evolution of certain feature descriptors for event detection. For instance, temporal evolutions of so-called visual descriptors such as "Lack of motion," "Fast pan," and "Fast zoom" were employed for soccer goal detection in [67], with the assumption that any interesting event affects two consecutive shots. In [60], the temporal relationships of the sub-events were studied to build event detection grammar. However, such setups are largely based on domain knowledge or human observations, which highly hinder the generalization and extensibility of the framework.

2. Secondly, the events of interest are often highly infrequent. Therefore, the classification techniques must deal with the class-imbalance (or called skewed data distribution) problem. The difficulties in learning to recognize rare events include: few examples to support the target class, the majority (i.e., nonevent) class dominating the learning process, etc.

In Section 6.1, a temporal segment analysis approach is proposed to address the above mentioned issues. However, its major focus is to explore the important temporal segments in characterizing events. In this section, a hierarchical temporal association mining approach is proposed to systematically address these issues.

In this approach, association rule mining and sequential pattern discovery are intelligently integrated to determine the temporal patterns for target events. In addition, an adaptive mechanism is adopted to update the minimum support and confidence threshold values by exploring the characteristics of the data patterns. Such an approach largely relaxes the dependence on domain knowledge or human efforts. Furthermore, the challenges posed by skewed data distribution are effectively tackled by exploring frequent patterns in the target class first and then validating them over the entire database. The mined temporal pattern is thereafter applied to further alleviate the class imbalance issue. As usual, soccer videos are used as the test bed.

6.2.1 Background

Association rules are an important type of knowledge representation revealing implicit relationships among the items present in a large number of transactions. Given $I = \{i_1, i_2, ..., i_n\}$ as the item space, a transaction is a set of items which is a subset of I. In the original market basket scenario, the items of a transaction represent items that were purchased concurrently by a user. An association rule is an implication of the form [$X \rightarrow Y$, *support*, *confidence*], where X and Y are sets of items (or itemsets) called antecedent and consequence of the rule with $X \subset I$, $Y \subset I$, and $X \bigcap Y = \emptyset$. The *support* of the rule is defined as the percentage of transactions that contain both X and Y among all transactions in the input data set; whereas the *confidence* shows the percentage of transactions that contain Y among transactions that contain X. The intended meaning of this rule is that the presence of X in a transaction implies the presence of Y in the same transaction with a certain probability. Therefore, traditional ARM aims to find frequent and strong association rules whose support and confidence values exceed the user-specified minimum *support* and minimum *confidence* thresholds.

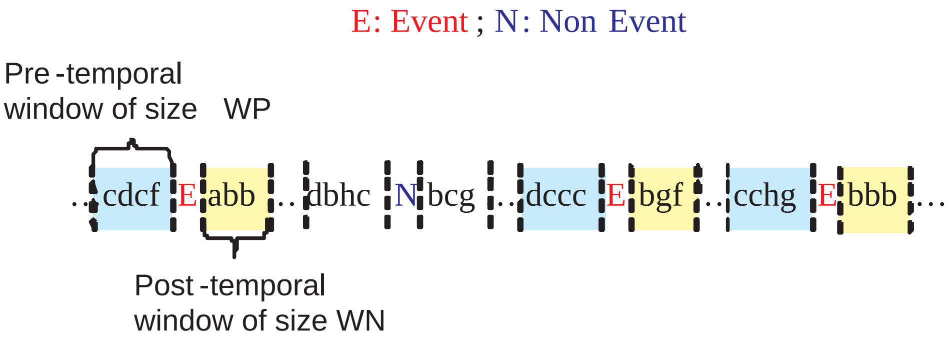

E: Event ; N: Non Event
Pre -temporal window of size WP
..cdcf E abb ..dbhc N bcg ..dccc E bgf ..cchg E bbb ...
Post -temporal window of size WN

Figure 6.6: An example video sequence.

Intuitively, the problem of finding temporal patterns can be converted as to find adjacent attributes (i.e., X) which have strong associations with (and thus characterize) the target event (i.e., Y), and thus ARM provides a possible solution. Here, assuming the analysis is conducted at the shot-level, the adjacent shots are deemed as the transaction and the attributes (items) can be the feature descriptors (low-, mid- or object-level extracted from different channels) or event types in the transaction. However, as discussed below, the problem of temporal pattern discovery for video event detection has its own unique characteristics, which differs greatly from the traditional ARM. Without loss of generalization, an event E is normally the result of previous actions (called pre-actions or AP) and might result in some effects (post-actions or AN). Given the example video sequence illustrated in Fig. 6.6, pre-transactions TP (such as $\{c, d, c, f\}$ and $\{d, c, c, c\}$) and post-transactions TN (such as $\{a, b, b\}$ and $\{b, b, b\}$) are defined as covered by the pre-temporal windows and post-temporal windows, respectively. The characters 'a,' 'b,' etc., denote the attributes of the adjacent shots. Note that if the feature descriptors are used as the attributes, certain discretization process should be conducted to create a set of discrete values to be used by ARM. A temporal context for target event E is thus composed of its corresponding pre-transaction and post-transaction, such as $< \{c, d, c, f\}\{a, b, b\} >$ and $< \{c, c, h, g\}\{b, b, b\} >$. The purpose of temporal association mining is thus to derive rules $< AP, AN > \rightarrow E$ that are frequent and strong, where

$AP \subset TP$ and $AN \subset TN$. Mainly, temporal pattern mining differs from the traditional ARM in two aspects.

- First, an itemset in traditional ARM contains only distinct items without considering the quantity of each item in the itemset. However, in event detection, it is indispensable that an event is characterized by not only the attribute type but also its occurrence frequency. For instance, in surveillance video, a car passing by a bank once is considered normal, whereas special attention might be required if the same car appears frequently within a temporal window around the building. In soccer video, several close views appearing in a temporal window might signal an interesting event, whereas one single close view is generally not a clear indicator. Therefore, a multiset concept is adopted which, as defined in mathematics, is a variation of a set that can contain the same item more than once. To our best knowledge, such an issue has not been addressed in the existing video event detection approaches. A slightly similar work was presented in [148], where ARM is applied to the temporal domain to facilitate event detection. However, it uses the traditional itemset concept. In addition, it searches the whole video to identify the frequent itemsets. Under the situation of rare event detection where the event class is largely under-represented, useful patterns are most likely overshadowed by the irrelevant itemsets.

- Second, in traditional ARM, the order of the items appearing in a transaction is considered as irrelevant. Therefore, transaction $\{a, b\}$ is treated the same as $\{b, a\}$. In fact, this is an essential feature adopted to address the issue of loose video structure. Specifically, the characteristic context information can occur at uneven inter-arrival times and display at different sequential orders as mentioned earlier. Therefore, given a reasonably small temporal window, it is preferable to ignore the appearance order of the attributes inside a pre-transaction or post-transaction.

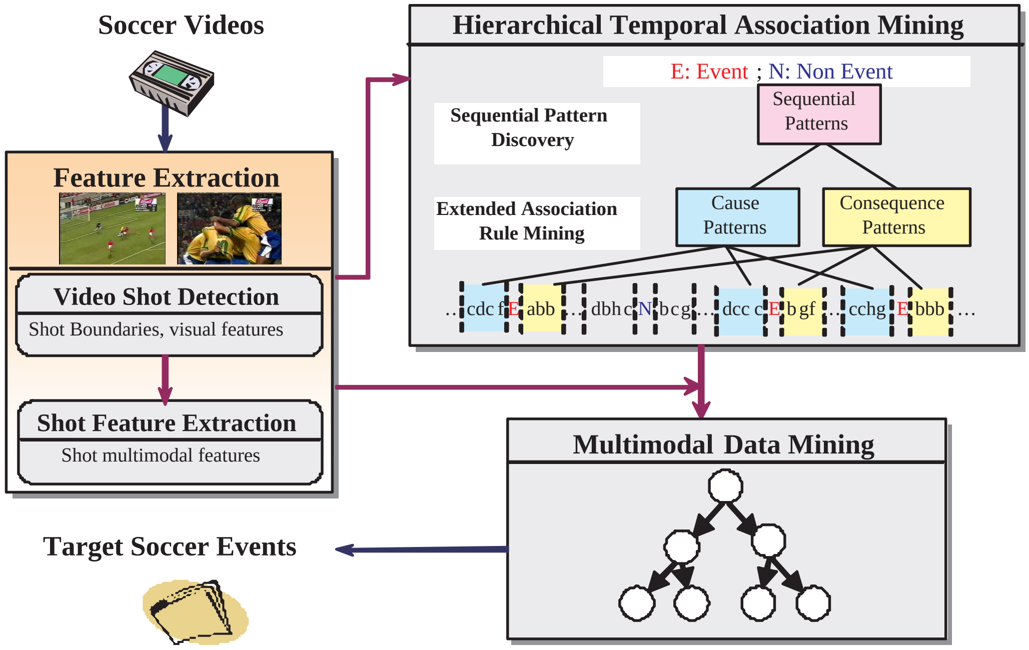


Figure 6.7: Hierarchical temporal mining for video event detection.

However, considering the rule $< AP, AN > \rightarrow E$, AP always occurs ahead of its corresponding AN, and the order between them is important in characterizing a target event. Therefore, in this stage, the idea of sequential pattern discovery [115] is adopted, where a sequence is defined as an ordered list of elements. In this study, each element is a multiset, that is, the sequence $< \{a, b\}\{c\} >$ is considered to be different from $< \{c\}\{a, b\} >$. In this paper, braces are used for multisets and angle brackets for sequences.

Fig. 6.7 shows the idea of using hierarchical temporal mining for video event detection. As compared to Fig. 5.7, a hierarchical temporal mining scheme is used to explore the knowledge assisted features, which will be detailed in the next section.

6.2.2 Hierarchical Temporal Association Mining

Since the target is to capture temporal patterns characterizing the contextual conditions around each target event, a hierarchical temporal association mining mechanism is proposed. As discussed earlier, due to the loose structure of videos, the attributes within the temporal windows (pre-temporal or post-temporal) have no orders. Meanwhile, the appearance frequency of the attributes is important in indicating the events. Hence, the proposed extended ARM algorithm is applied to find pre-actions AP and post-actions AN (called "Extended ARM" in Fig. 6.7), and then sequential pattern discovery is utilized where AP and AN are considered as the elements in a sequence (called "Sequential Patterns" in Fig. 6.7). Thereafter, the temporal rules are derived from the frequent and strong patterns. The approach is first presented with the predefined minimum support and confidence thresholds, and an adaptive updating mechanism is introduced to define them automatically.

Let $D_v = \{V_i\}$ be the training video database and NF be the number of attributes in the database, where $V_i (i = 1, ..., N_v)$ is a video clip and N_v is the cardinality of D_v, we have the following definitions.

Definition 6.1. A video sequence V_i is an ordered collection of units $V_i =< V_{i1}, V_{i2}, ..., >$, where each unit $V_{ij} (j = 1, ..., n_i)$ is a 3-tuple $V_{ij} = (F_{ij}, s_{ij}, C_{ij})$. Here, n_i is the number of units in V_i, $F_{ij} = \{F_{ijk}\}$ indicates the set of unit attributes $(k = 1, ..., N_F)$, s_{ij} denotes its associated unit number, and $C_{ij} = \{yes, no\}$ is the class label showing the eventness of the unit.

In this study, the unit is defined at the shot level and the unit attribute, as mentioned earlier, can be the feature descriptor or event type of the shot. As usual, the task is to find all frequent and strong patterns from the transactions given the target event E. Therefore, the pre-transactions (TP) and post-transactions (TN) need to be constructed.

Definition 6.2. Given a unit $V_{ij} (j = WP+1, ..., n_i - WN)$, the pre-temporal window size WP and post-temporal window size WN, its associated TP_{ij} and TN_{ij} are defined as $TP_{ij} = \{F_{ip}\}$ $(p = j - WP, ..., j - 1)$ and $TN_{ij} = \{F_{iq}\}$ $(q = j + 1, ..., j + WN)$.

Frequent patterns

This proceeds by first finding all frequent patterns. Different from traditional ARM, to alleviate the problem of class imbalance problem, the frequent patterns are searched for the minority class only. In other words, in counting the frequent patterns and calculating the support values, only those $TP_E = \{TP_{ij}\}$ and $TN_E = \{TN_{ij}\}$ will be checked where $C_{ij} =$ '*yes*.' As shown in Fig. 6.6, the multisets $\{d, b, h, c\}$ and $\{b, c, g\}$ around the nonevent N will not be checked in this step. Then the discrimination power of the patterns is validated against the nonevent class.

In order to mine the frequent pre-actions and post-actions, the itemMultiset (the counterpart of itemset in traditional ARM) is defined.

Definition 6.3. An itemMultiset T is a combination of unit attributes. T matches the characterization of an event in window WP or WN if T is the subset of TP_{ij} or TN_{ij} where $C_{ij} =$ '*yes*.'

For example, if a post-temporal window with size WN for an event E (see Fig. 6.6) contains unit attributes $\{a, b, b\}$, then $T = \{b, b\}$ is called a match of the characterization of event E, whereas $T = \{a, a\}$ is not. Consequently, the traditional support and confidence thresholds are revised as follows.

Definition 6.4. An itemMultiset T has support s in D_v if s% of all $TP_E = \{TP_{ij}\}$ (or $TN_E = \{TN_{ij}\}$) for target event E are matched by T. T is frequent if s exceeds the predefined *min_sup*.

Mathematically, support is defined as

$$Support = Count(T, TP_E)/|TP_E| \tag{6.9}$$

or

$$Support = Count(T, TN_E)/|TN_E| \tag{6.10}$$

From the equations, it can be seen that the definition of support is not simply an extension of the one used in traditional ARM. It is restricted to $TP_E = \{TP_{ij}\}$ or $TN_E = \{TN_{ij}\}$ which are associated with the target events (i.e., C_{ij} = 'yes'). An itemMultiset which appears in D_v periodically might not be considered as frequent if it fails to be covered by these TP_E or TN_E. The pseudo code for finding frequent itemMultisets is listed in Table 6.2. The general idea is to maintain in memory, for each target event, all the units within its associated TP_{ij} and TN_{ij}, which are then stored in B_p and B_n (steps 1 to 19), and extended Apriori algorithm is applied to find the frequent pre-actions and post-actions from B_p and B_n (steps 20 to 21).

Table 6.2: Logic to find all frequent unit patterns.

```
Algorithm 1: Finding Frequent Patterns
Input: video database $D_v$, pre-temporal window size $WP$, post-temporal window
size $WN$, minimum support $min\_sup$, target-event type $E$
Output: frequent actions $AP$, $AN$
FrequentActions($D_v$, $WP$, $WN$, $min\_sup$, $E$)
1. $B_p = \emptyset$; $T = \emptyset$; $B_n = \emptyset$
2. for each video sequence $V_i \in D_v$
3.    for each unit $V_{ij} = (F_{ij}, s_{ij}, C_{ij}) \in V_i$
4.       for each unit $V_{ik} = (F_{ik}, s_{ik}, C_{ik}) \in T$
5.          if $(s_{ij} - s_{ik}) > WP$
6.             Remove $V_{ik}$ from $T$
7.          endif
8.       endfor
9.       if $V_{ij}$ is a target event // i.e., $C_{ij}$ = 'yes'
10.          $B_p = B_p \cup \{F_{ik} | (F_{ik}, ) \in T\}$
11.          $PS = s_{ij} + 1$
12.          while $(PS - s_{ij}) < WN$
13.             $B_n = B_n \cup \{F_{ik} | s_{ik} = PS\}$
14.             $PS$ is set to its next shot until it is the end of $V_i$
15.          endwhile
16.       endif
17.       $T = T \cup V_{ij}$
18.    endfor
19. endfor
20. Use extended Apriori over $B_p$ to find $AP$ with $min\_sup$
21. Use extended Apriori over $B_n$ to find $AN$ with $min\_sup$
```

Table 6.3: The procedure of extended A-priori algorithm.

1. Construct 1-itemMultisets. Count their supports and obtain the set of all frequent 1-itemMultisets as traditional A-priori algorithm

2. A pair of frequent k-itemMultisets are merged to produce a candidate $(k+1)$-itemMultisets. The merges are conducted in two steps:

2.1. A pair of frequent k-itemMultisets are merged if their first $(k-1)$ items are identical and

2.2 A frequent k-itemMultisets can be merged with itself only if all the elements in the multiset are with the same value

3. The supports are counted and the frequent itemMultisets are obtained as the traditional A-priori algorithm. Go to step 2.

4. The algorithm terminates when no further merge can be conducted.

The procedure of the extended Apriori algorithm is shown in Table 6.3, which will be explained by an example. Since in the transactions (TP or TN) and itemMultisets the existence of duplicated elements is allowed, each unit attribute needs to be considered as a distinct element even though some attributes might have the same values, except for the construction of 1-itemMultisets. The frequent pre-patterns and post-patterns, obtained by using the proposed extended Apriori algorithm upon the example video sequence shown in Fig. 6.6, are listed in Tables 6.4 and 6.5, respectively.

Here, it is assumed that the minimum support count is set to 2 and the frequent actions are highlighted in yellow. Since the ordering of the units and the inter-arrival times between the units and target events within each time window is considered to be irrelevant in finding the frequent pre- and post-patterns, for the sake of simplicity, all the units inside the transactions and itemMultisets are sorted in the algorithm. The computational cost for such procedures is minimal because the transactions are constructed only for minority class and the number of elements in such transactions is small. Without loss of generality, the window size is reasonably small since only the temporally adjacent shots have strong association with the target events.

As mentioned earlier, the ordering between pre-actions AP and post-actions AN needs to be observed, and thus the idea of sequential pattern discovery is adopted (omitting the

Table 6.4: Frequent pre-actions.

1	**count**	**frequent**	**2**	**count**	**frequent**	**3**	**count**	**frequent**
$\{c\}$	3	Yes	$\{c, c\}$	3	Yes	$\{c, c, c\}$	1	No
$\{d\}$	2	Yes	$\{c, d\}$	2	Yes	$\{c, c, d\}$	2	Yes
$\{f\}$	1	No	$\{d, d\}$	0	No	-	-	-
$\{g\}$	1	No	-	-	-	-	-	-
$\{h\}$	1	No	-	-	-	-	-	-

Table 6.5: Frequent post-actions.

1	**count**	**frequent**	**2**	**count**	**frequent**	**3**	**count**	**frequent**
$\{a\}$	1	No	$\{b, b\}$	2	Yes	$\{b, b, b\}$	1	No
$\{b\}$	3	Yes	-	-	-	-	-	-
$\{g\}$	1	No	-	-	-	-	-	-
$\{f\}$	1	No	-	-	-	-	-	-

detailed algorithm here). However, it is worth noting that instead of scanning TP and TN to explore the frequent sequential patterns, the Apriori like principle can be applied to simplify the process, which states that for a particular sequence to be frequent, its element(s) must be frequent as well. For instance, given the examples shown in Fig. 6.6 and frequent pre- and post-actions listed above, respectively, it can be known that sequence $< \{a\}\{d\} >$ is not frequent since its pre-action element $< \{a\} >$ is not frequent. Therefore, the frequent sequential patterns can be constructed upon the frequent AP and AN. It is legal to have null pre-action or post-action in a sequential pattern (e.g., $< \{\}\{b, b\} >$ or $< \{c, c, d\}\{\} >$).

After creating the 1-itemMutlisets, the corresponding sequential patterns can be extracted. Then when another pass is made over the transactions to find frequent 2-itemMultisets, the support of the constructed sequential pattern can be counted as well. The procedure terminates until no more frequent (k+1)-itemMultisets can be identified.

Strong patterns

To validate that these patterns effectively characterize the event of interest, a restrict solution is to adopt the traditional association measure called confidence, where a similar idea presented in [121] can be adopted. The general idea is to count the number of times each of the patterns occurs outside the windows of the target events.

Definition 6.5. A sequential pattern P has confident c in D_v if c% of all transactions matched by T are associated with the target event. P is strong if c exceeds min_conf.

Intuitively, inputs of a set of transactions are checked, which correspond to all $TP_N = \{TP_{ij}\}$ and $TN_N = \{TN_{ij}\}$ with C_{ij} = 'no.' In fact, such lists can be obtained in algorithm 1 when scanning through the unit sequence and storing them in $B_p^{'}$ and $B_n^{'}$, respectively. Let x_1 and x_2 be the counts when the pattern T is matched in B and $B^{'}$. Here $B = \{b_1, b_2, ..., b_n\}$ is constructed by linking $B_p = \{b_{p1}, b_{p2}, ..., b_{pn}\}$ and $B_n = \{b_{n1}, b_{n2}, ..., b_{nn}\}$, where $b_i =< b_{pi}, b_{ni} >$. Similarly, $B^{'}$ can be constructed by $B_p^{'}$ and $B_n^{'}$. The confidence of P is defined as follows.

$$confidence(P, B, B^{'}) = x_1/(x_1 + x_2) \tag{6.11}$$

This metric is thus applied to compare with min_conf and to validate whether the temporal patterns are strong.

6.2.3 Temporal rules

Once the frequent and strong temporal patterns are obtained, temporal rules can be built to facilitate the event detection. The principle is defined as follows.

Definition 6.6. Given two patterns, P_i and P_j, $P_i \succ P_j$ (also called P_i has a higher rank than P_j) if

1. The confidence of P_i is greater than that of P_j, or

2. Their confidences are the same, but the support of P_i is greater than that of P_j, or

3. Both the confidences and supports of P_i and P_j are the same, but P_i is more specific than P_j (i.e., P_j is a subsequence of P_j).

The rules are in the form of $P_i \rightarrow E$ (targeted event). Let R be the set of generated rules and D_v be the training database. The basic idea is to choose a set of high ranked rules in R to cover all the target events in D_v. Such temporal rules are applied in the data pruning process to generate a candidate event set and to alleviate class imbalance problem in the data classification stage.

Adaptive metrics updating mechanism

The performance of the proposed approach is partially related to four parameters: WP, WN, min_sup and min_conf. Among them, WP and WN can be determined relatively straightforward as generally only the temporally adjacent shots have strong association with the target events. Therefore, they can be set to any reasonably small values such as 3 or 4. In addition, as discussed in Section 6.1, an advanced approach was proposed to identify the significant temporal window with regard to the target event, which can be incorporated into this framework to define the window size. Therefore, in this section, an adaptive metrics updating mechanism is proposed to define min_sup and min_conf in an iterative manner.

The richness of the generated patterns is partially dependent on min_sup, which in most existing works is defined manually based on domain knowledge. However, given a training database, it is infeasible to expect users to possess knowledge of the complete characteristics of the training set. Therefore, the proposed approach addresses this issue by refining the support threshold $SupTH_{k+1}$ iteratively based on the statistical analysis of the frequent patterns obtained using threshold $SupTH_k$. Given k^{th} threshold $SupTH_k$, let R_k be the number of attributes in the largest frequent itemMultisets, we have $Sup_{kr} = \{supportsofallr - itemMultisets\}$, where $r = 1, ..., R_k$. Equations 6.12 to 6.14 define min_sup.

$$diff(r) = mean(Sup_{kr}) - mean(Sup_{kr+1}), r = 1, ..., R_{k-1} \tag{6.12}$$

$$r_k = argmax_r(diff(r)) \tag{6.13}$$

$$\begin{cases} if diff(r_k) > R_k/2 & SupTH_{k+1} = (mean(Sup_{kr}) - mean(Sup_{kr+1}) \\ else & min\_sup = SupTH_k \end{cases} \tag{6.14}$$

where r = 1, ..., R_k - 1.

The idea is that the learned frequent patterns in the previous round can help reveal certain knowledge regarding the training data set and thus help refine the support threshold intelligently. Specifically, the biggest fluctuation is studied between the supports of two adjacent itemMultisets. Since $(r+1)$-itemMultisets are rooted from r-itemMultisets, if the difference is greater than $R_k/2$, the support threshold is adjusted to avoid the possible over-fitting issue and improve framework efficiency. Note that the initial support threshold $SupTH_0$ can be set to a reasonably small value.

For the confidence threshold, a similar criterion is adopted to examine the biggest difference between two adjacent sequential patterns with the condition that the generated rules in R should be able to cover all target events in D_v. In other words, if the newly defined confidence threshold $ConTH_{k+1}$ causes the missing of target events in D_v, $ConTH_k$ is chosen as min_conf.

6.2.4 Experiments

To testify the effectiveness of the proposed temporal association mining approach, the same experimental data set used in both Chapter 5 and Section 6.1 was adopted. A set of temporal association rules were generated following the procedure addressed in Sections 6.2.2 and 6.2.3. To apply the temporal association mining, in current imple-

Table 6.6: Performance of goal event detection using temporal association mining.

	# of goals	**Identified**	**Missed**	**Misidentified**	**Recall**	**Precision**
Test 1	11	10	1	2	90.9%	83.3%
Test 2	14	12	2	3	85.7%	80.0%
Test 3	12	11	1	2	91.7%	84.6%
Test 4	11	11	0	2	100.0%	84.6%
Test 5	12	11	1	1	91.7%	91.7%
				Average	92.0%	84.8%

mentation, a discretization process is performed first to convert the continuous feature values into nominal values. In future work, fuzzy logic might be applied in this step to further improve the performance. The constructed rules were thus applied as a data reduction step to alleviate the class imbalance issue. Finally, the decision tree logic was applied upon the 'cleaned' data set for event detection. Similarly, a 5-fold cross validation scheme was adopted and the same metrics, recall and precision, were used to evaluate the framework performance. Table 6.6 shows the experimental results. As can be seen, the performance is improved in comparison to the results shown in Table 6.1 as the temporal association mining offers an intelligent approach to not only capture but also represent the characteristic temporal patterns.

The precision and recall values were computed for all the testing data sets in these five groups (denoted as Test 1, Test 2, etc.) to evaluate the performance of the proposed framework. As shown in Table 5, the "Missed" column indicates a false negative, which means that the goal events are misclassified as nongoal events; whereas the "Misiden" column represents a false positive, i.e., the nongoal events are identified as goal events.

From the above results, it can be clearly seen that the performance is quite promising in the sense that the average recall and precision values reach 96.5% and 84.1%, respectively. In addition, the performance across different testing data sets is greatly consistent. Furthermore, the dependency on predefined domain knowledge is largely relaxed since an automatic temporal association mining process is adopted in the framework to discover, represent, and apply the characteristic event temporal patterns. As a result, the framework possesses a greater potential to be applied to different domains.

6.2.5 Conclusions

As discussed in Section 2.2.3, currently high level indexing techniques are primarily designed from the perspective of manual indexing or annotation as automatic high-level video content understanding is still infeasible for general videos. With the ultimate goal of developing a general and flexible framework which can be applied to different domains with minor extra effort, a key aspect is to largely relax the reliance on domain knowledge or *a priori* information. In response to such demand, in this section, an innovative temporal association mining approach is proposed to effectively capture and represent the characteristic context information for interesting events. Compared to most existing works, the dependence on domain knowledge is largely relaxed with the assistance of the automatic knowledge discovery method. In addition, different from the approach discussed in Section 6.1, this framework offers a systematic principle to represent the significant context information and takes into consideration the special challenges posed by the class imbalance issue. This approach is thus an initial yet critical step in the continuous efforts in automating the high-level indexing process. The effectiveness of this framework is fully demonstrated by the experimental results.

CHAPTER 7

Conclusions and Future Work

In this dissertation, a knowledge-assisted data management and retrieval framework is proposed for Multimedia Database Management Systems (MMDBMSs). The main focus of this work is to address three essential challenges: semantic gap, perception subjectivity and data management. Taking image and video as the test beds, a variety of techniques are proposed to address these challenges in three main aspects of a MMDBMS: multimedia data representation, indexing and retrieval.

In terms of image data representation, low-level features, such as color and texture, are extracted from images. In addition, to capture the salient object information in the images, an unsupervised segmentation technique called WavSeg is adopted to decompose the images into homogeneous regions, where the object-level features are captured correspondingly. Although a set of low-level and object-level features are captured by a number of advanced techniques, they alone are inadequate to model the comprehensive image content (semantic meanings). Therefore, a semantic network approach is proposed to model the semi-semantic representation of the images in the image database. The semantic network adopts the Markov Model Mediator concept and stochastically models the affinity relationships among the images based on the accumulated feedback logs in the database. As each feedback contains valuable semantic information with respect to the similarity of the images, by probabilistic reasoning, the high-level knowledge from general users' viewpoints is not only captured, but also systematically modeled by the semantic network to bridge the semantic gap.

To construct the video data representation, in this work videos are first decomposed into a set of meaningful and manageable units, i.e., shots for analyzing. Shot-level multimodal features (visual and audio features) are then obtained by averaging frame features across the entire shot. Alternatively, key frame(s) can be extracted to serve as a representation of the corresponding shots and its content is processed. Since each frame is in fact a static image, some of the techniques, such as WavSeg algorithm, color/texture feature extraction, adopted for image content analysis are also applied for shot boundary detection and visual feature extraction. Since video streams are with complicated content form and inherent temporal dimensions, a mid-level representation and temporal knowledge discovery approaches are adopted to bridge the semantic gap. As discussed in Section 5.2.2, the advantage of introducing mid-level representation is that it offers a reasonable tradeoff between the computational requirements and resulting semantics. The effectiveness of mid-level representation is fully demonstrated by the experiments. However, it requires certain levels of domain knowledge and human effort. To relax such dependency, two advanced knowledge discovery approaches are proposed with the aim to automatically capture the characteristic context information to assist the video semantic content analysis.

The various levels of media data representations result in a multi-level indexing scheme to accommodate different kinds of query and retrieval requirements. In particular, for video database management, a data classification mechanism is presented for high-level video event detection and annotation with the assistance of both low-level features and mid-level or knowledge-assisted data representations.

To serve for user's specific query interests (i.e., perception subjectivity) and at the same time ensure a fast convergence process, the MMM mechanism and RF technique are integrated seamlessly to capture users' perception in the image level. In addition, the MMIR framework is proposed to effectively model users' perception at both the image and object-level based on users' interactions. Furthermore, the MMM mechanism is extended to enable image database clustering and cluster-based image retrieval to support efficient image retrieval in a distributed environment.

On the basis of current research results, the future work is proposed accordingly as listed below.

1. Integration of multimodal features: In the current video mining framework, the integration of different modalities is conducted by manually analyzing the temporal evolution of the features within each modality and the temporal relationships between different modalities. More specifically, the audio and visual features are aligned in the shot-level and it in many cases this might not be an optimal solution. In future work, the modeling of each modality will be conducted by using statistical models or temporal association rule mining scheme such that different modalities can be integrated and temporal constraints can be accommodated.

2. The automatic temporal knowledge discovery provides great potential to facilitate high-level video analysis, indexing and annotation as they aim to relax the framework's dependence on domain knowledge or human effort. However, further research effort is required to enhance these approaches. For instance,

    - In the current temporal association mining algorithm, the size of temporal window on which the algorithm is applied is not yet well-defined. A simple assumption is adopted that for a temporal pattern to be significant in characterizing a certain event, it should be found in its adjacent temporal segments and the size is thus set to 5 (i.e., the temporal segment contains 5 consecutive shots). Such a setting is rather ad hoc and is not dynamic enough to model different events in various applications. In future work, the effects caused by various window sizes should be first studied and a systematic method should be proposed to determine the window size intelligently. In fact, the temporal segment analysis algorithm targets deciding the size and location of the temporal segment. Therefore, one possible solution is to integrate the temporal segment analysis approach with the temporal association mining framework. Alternatively, the window size can be defined by a function associated with the performance metrics, such as support and confidence values.

- The temporal association mining algorithm should be performed upon the nominal attributes. For continuous values, a discretization process should be applied beforehand. Currently, this process is conducted empirically. In future work, fuzzy discretization or other discretization approaches might be introduced in this step to reduce the information loss and boost the framework performance.

- Effective spatio-temporal indexing for video database remains an open issue. In this study, the proposed temporal segment analysis and temporal association mining algorithm possess the capabilities of capturing the characteristic temporal segment and important context information for a specific event unit. Such information is essential not only for event detection but for temporal indexing, as basically it represents the temporal evolution of the video activities. Future research can be conducted to construct a feasible mechanism to utilize the temporal analysis results in temporal indexing.

3. In terms of video event detection, the current classification algorithm aims to minimize the expected number of errors with the assumption that the costs of different misclassification errors are identical. However, in many real application domains such as medical diagnosis, surveillance videos or even sports videos, the event class is usually rare and the cost of missing a target event (false negative) is generally greater than including a nonevent (false positive). In such domains, classifier learning methods that do not take misclassification costs into account might not perform well. For instance, in rare event detection, the influence of rare events will be overshadowed by the majority class and the classification model is built in favor of the

majority class. Therefore, cost-sensitive classification approaches can be adopted which perform the data classification with non-uniform costs. A possible solution is to define the cost matrix where the cost of a false negative is set to be higher than that of a false positive. Essentially, the goal is to build a classifier to minimize the expected misclassification costs rather than to minimize the expected number of misclassification errors.

4. In general, a large set of attributes are extracted to represent the media content. However, such high-dimensional data representation poses great challenges towards media data management. In fact, various attributes might be correlated in the sense that they contain redundancy information. In addition, some features contain noisy information which actually deteriorates the overall performance. Moreover, the discrimination between classes becomes much more difficult with a high dimensional feature set because the training samples are likely scattered in a high-dimensional space. Therefore, an automatic feature selection scheme is of great importance to improve both the effectiveness and efficiency of a MMDBMS.

5. Future research efforts will also be directed to develop a better interaction scheme and faster converging process to alleviate the manual effort in media retrieval. Generally, for an interactive CBR system, the query performance is achieved at the cost of huge human effort. Take the relevance feedback system as an example. Normally, users are asked to go through 3 to 4 iterations to provide their feedback (positive, negative, or even the level of relativity in some approaches) for tens of images in each iteration. It can be expected that the level of manual effort required for image retrieval will be one of the most important factors that determine the potential and popularity of the CBR system in the real application domains. In this proposal, a semantic network is proposed to accumulate and analyze the historical feedback to improve long-term system performance and to speed up the converging

process. However, currently the feedback information is not fully utilized as only positive feedback is analyzed in constructing the semantic network. This work can be further extended to probabilistic learning of not only the positive and negative feedback, but also the level of relativity (if provided).

6. With the prosperity of the Internet, the Web has become a huge repository for various media data and image retrieval from the Web has attracted increasing attention. Some well-known search engines such as Google and Yahoo! generate search results by using automated web crawlers, such as spider and robot. However, general search engines often provide inaccurate search results since they are designed to be keyword-based retrieval, which as discussed in Section 2.1.3 has large limitations. A solution to address this issue is to construct the web crawler by using a content-based image retrieval mechanism, which improves the rudimentary searching results provided by the general Internet search tools. For this exciting new application domain, many techniques addressed in this proposal can be applied but need further adjustment or customization. For instance, although low-level feature extraction in general image databases has certain efficiency requirements, it is becoming even more critical for Web image searching. In addition, a semantic network constructed for general image databases effectively bridges the semantic gap by stochastically analyzing the accumulated feedback log. Intuitively, this mechanism will be of great help for Web image searching as well. However, a problem exists as how to collect and accumulate the feedback logs. It is also possible that the semantic network needs to be constructed by using different information sources.

Each of the topics mentioned above is of great importance for a successful MMDBMS and will be addressed by leveraging the current research framework.

[1] S. Ardizzoni, I. Bartolini, and M. Patella, “Windsurf: Region-Based Image Retrieval Using Wavelets,” in *Proceedings of the 10th International Workshop on Database and Expert Systems Applications (DEXA)*, pp. 167-173, 1999.

[2] N. Babaguchi, Y. Kawai, and T. Kitahashi, “Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration,” *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 68-75, 2002.

[3] R. Brunelli, O. Mich, and C.M. Modena, “A Survey on the Automatic Indexing of Video Data,” *Journal of Visual Communication and Image Representation*, vol. 10, pp. 78-112, 1999.

[4] C. Carson, S. Belongie, H. Greenspan, and J. Malik, “Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 10261038, 2002.

[5] V. Castelli and L. D. Bergman, Image Databases: Search and Retrieval of Digital Imagery. New York John Wiley & Sons, Inc. ISBN: 0471321168.

[6] S.-F. Chang, W. Chen, et al., “A Fully Automatic Content-Based Video Search Engine Supporting Multi-Object Spatio-temporal Queries,” *IEEE Transactions on Circuits and Systems for Video Technology*, Special Issue on Image and Video Processing for Interactive Multimedia, vol. 8, no. 5, pp. 602-615, 1998.

[7] M. Chen and S.-C. Chen, “MMIR: An Advanced Content-based Image Retrieval System using a Hierarchical Learning Framework,” accepted for publication, Edited by J. Tsai and D. Zhang, *Advances in Machine Learning Application in Software Engineering*, Idea Group Publishing.

[8] M. Chen, S.-C. Chen, et al., “Semantic Event Detection via Temporal Analysis and Multimodal Data Mining,” *IEEE Signal Processing Magazine*, Special Issue on Semantic Retrieval of Multimedia, vol. 23, no. 2, pp. 38-46, 2006.

[9] M. Chen, S.-C. Chen, M.-L. Shyu, and C. Zhang, “Video Event Mining via Multimodal Content Analysis and Classification,” accepted for publication, Edited by V. A. Petrushin and L. Khan, *Multimedia Data mining and Knowledge Discovery*, Springer-Verlag.

[10] S.-C. Chen, M. Chen, C. Zhang, and M.-L. Shyu, “Exciting Event Detection using Multi-level Multimodal Descriptors and Data Classification,” in *Proceedings of IEEE International Symposium on Multimedia*, pp. 193-200, 2006.

[11] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, “A Multimodal Data Mining Framework for Soccer Goal Detection Based on Decision Tree Logic,” *International Journal of Computer Applications in Technology*, vol. 27, no. 4, 2006.

[12] L. Chen and M. T. Ozsu, "Modeling of Video Objects in a Video Database," in *Proceedings of IEEE International Conference on Multimedia*, pp. 217-221, 2002.

[13] S.-C. Chen, S. H. Rubin, M.-L. Shyu, and C. Zhang, "A Dynamic User Concept Pattern Learning Framework for Content-Based Image Retrieval," *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, vol. 36, no. 6, pp. 772-783, 2006.

[14] S.-C. Chen, M.-L. Shyu, M. Chen, and C. Zhang, "A Decision Tree-based Multimodal Data Mining Framework for Soccer Goal Detection," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 265-268, 2004.

[15] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Video Scene Change Detection Method Using Unsupervised Segmentation and Object Tracking," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 57-60, 2001.

[16] S.-C. Chen, M.-L. Shyu, C. Zhang, and R. L. Kashyap, "Identifying Overlapped Objects for Video Indexing and Modeling in Multimedia Database Systems," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 715-734, 2001.

[17] S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative Shot Boundary Detection for Video Indexing," edited by Sagarmay Deb, *Video Data Management and Information Retrieval*. Idea Group Publishing, ISBN: 1-59140546-7; pp. 217-236, 2005.

[18] S.-C. Chen, M.-L. Shyu, N. Zhao, and C. Zhang, "Component-Based Design and Integration of a Distributed Multimedia Management System," in *Proceedings of the 2003 IEEE International Conference on Information Reuse and Integration*, pp. 485-492, 2003.

[19] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "An Indexing and Searching Structure for Multimedia Database Systems," in *Proceedings of the IS&T/SPIE International Conference on Storage and Retrieval for Media Databases*, pp. 262-270, 2000.

[20] Y. Chen and J. Z. Wang, "A Region-based fuzzy feature matching approach to content-based image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1252-1267, 2002.

[21] Y. Chen, and J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions," *Journal of Machine Learning Research*, vol. 5, pp. 913-939, 2004.

[22] X. Chen, C. Zhang, S.-C. Chen, and M. Chen, "A Latent Semantic Indexing Based Method for Solving Multiple Instance Learning Problem in Region-Based Image Retrieval," in *Proceedings of IEEE International Symposium on Multimedia*, pp. 37-44, 2005.

[23] H. D. Cheng and Y. Sun, "A Hierarchical Approach to Color Image Segmentation Using Homogeneity," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2071-2082, 2001.

[24] P. Ciaccia, M. Patella, and P. Zezula, "M-tree: An Efficient Access Method for Similarity Search in Metric Spaces," in *Proceedings of the 23rd VLDB conference*, pp. 426-435, 1997.

[25] M. Cooper, J. Foote, and A. Girgensohn, "Temporal Event Clustering for Digital Photo Collections," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, pp. 364-373, 2003.

[26] I. J. Cox, M. L. Miller, et al., "The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments," *IEEE Transaction on Image Processing*, vol. 9, no. 1, pp. 20-37, 2000.

[27] J. D. Courtney, "Automatic Video Indexing via Object Motion Analysis," *Pattern Recognition*, vol. 30, no. 4, pp. 607-625, 1997.

[28] S. Dagtas and M. Abdel-Mottaleb, "Extraction of TV Highlights Using Multimedia Features," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, pp. 91-96, 2001.

[29] S. Dao, Q. Yang, and A. Vellaikal, "MB+-tree: An Index Structure for Content-Based Retrieval," in Chapter 11 of *Multimedia Database Systems: Design and Implementation Strategies*, MA: Kluwer, 1996.

[30] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the Multiple-Instance Problem with Axis-Parallel Rectangles." *Artificial Intelligence*, vol. 89, pp. 31-71, 1997.

[31] L.-Y. Duan, et al., "A Mid-level Representative Framework for Semantic Sports Video Analysis," in *Proceedings of ACM International Conference on Multimedia*, pp. 33-44, 2003.

[32] A. Ekin, A. M. Tekalp, R. Mehrotra, "Automatic Soccer Video Analysis and Summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796-807, 2003.

[33] W. E. Farag and H. Abdel-Wahab, "Video Content-Based Retrieval Techniques," Edited by Sagarmay Deb, *Multimedia Systems and Content-Based Retrieval*, Idea Group Publishing, pp. 114-154, 2005, ISBN: 1-59140546-7.

[34] X. Feng and H. Huang, "A Fuzzy-Set-Based Reconstructed Phase Space Method for Identification of Temporal Patterns in Complex Time Series," *IEEE Transaction on Knowledge and Data Engineering*, vol. 17, no. 5, pp. 601-613, 2005.

[35] M. Flickner, et al., "Query By Image and Video Content: The QBIC System," *IEEE Computer*, vol. 28, no. 9, pp. 23-32, 1995.

[36] O. Frank and D. Strauss, "Markov Graphs," *Journal of the American Statistical Association*, vol. 81, pp. 832-842, 1986.

[37] E. B. Goldstein, Sensation and Perception. Brooks/Cole.

[38] Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi, "Automatic Parsing of TV Soccer Programs," in *Proceedings of IEEE Multimedia Computing and Systems*, 1995.

[39] Google Images, http://images.google.com/

[40] A. Gupta and R. Jain, "Visual Information Retrieval," *Communications of the ACM*, vol. 40, no. 5, pp. 71-79, 1997.

[41] M. Haas, M. S. Lew, and D. P. Huijsmans, "A New Method for Key Frame Based Video Content Representation," edited by A. Smeulders and R. Jain, *Image Databases and Multimedia Search* , World Scientific., pp. 191-200, 1997.

[42] J. Hafner, H. S. Sawhney, W. Equitz, M.Flickner and W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729-736, July, 1995.

[43] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107-145, 2001.

[44] M. Han, W. Hua, W. Xu, and Y. Gong, "An Integrated Baseball Digest System Using Maximum Entropy Method," in *Proceedings of the 10th ACM International Conference on Multimedia*, pp. 347-350, 2002.

[45] J. Han and M. Kamber, *Data Mining C Concepts and Techniques*, Morgan Kaufmann, ISBN: 1-55860-489-8, 2001

[46] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 90-105, 2002.

[47] X. He, O. King, W.-Y. Ma, M. Li, and H. J. Zhang, "Learning a Semantic Space from User's Relevance Feedback for Image Retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 1, pp. 39-49, 2003.

[48] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang, "Mean Version Space: A New Active Learning Method for Content-based Image Retrieval," in *Proceedings of ACM International Conference on Multimedia*, pp. 15-22, 2004.

[49] R. Heisterkamp and J. Peng, "Kernel VA-Files for Relevance Feedback Retrieval," in *Proceedings of the First ACM International Workshop on Multimedia Databases*, pp. 48-54, 2003.

[50] C. H. Hoi and M. R. Lyu, "A Novel Log-based Relevance Feedback Technique in Content-Based Image Retrieval," in *proceedings of ACM International Conference on Multimedia*, pp. 24-31, 2004.

[51] X. Huang, , S.-C. Chen, and M.-L. Shyu, "Incorporating Real-Valued Multiple Instance Learning into Relevance Feedback for Image Retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 321-324, 2003.

[52] T.-H. Hwang and D.-S. Jeong, "Detection of Video Scene Breaks Using Directional Information in DCT Domain," in *Proceedings of the 10th International Conference on Image Analysis and Processing*, pp. 887-892, 1999.

[53] S. Intille and A. Bobick, "Recognizing Planned, Multi-person Action," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 414-445, Mar. 2001.

[54] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: Query Databases through Multiple Examples," in *Proceedings of the 24th VLDB Conference*, pp. 218-227, 1998.

[55] X. Jin and J. C. French, "Improving Image Retrieval Effectiveness via Multiple Queries," in *Proceedings of ACM International Workshop on Multimedia Database*, pp. 86-93, 2003.

[56] F. Jing, M. Li, H. J. Zhang, and B. Zhang, "An Effective Region-Based Image Retrieval Framework," in *Proceedings of ACM International Conference on Multimedia*, pp. 456-465, 2002.

[57] T. Joachims, "Making Large-Scale SVM Learning Practical," Edited by B. Scholkopf, C. Burges, and A. Smola, *Advances in Kernel Methods-Support Vector Learning*, MIT-Press, 1999.

[58] J. D. Jobson, Applied Multivariate Data Analysis Volume II: Categorical and Multivariate Methods. Springer-Verlag Inc., NY, 1992.

[59] L. M. Kaplan, et al., "Fast Texture Database Retrieval Using Extended Fractal Features," in *Proceedings of IS&T/SPIE Conference on Storage and Retrieval for Media Databases*, pp. 162-173, 1998.

[60] Y.-L. Kang, J.-H. Lim, et al. "Soccer Video Event Detection with Visual Keywords," in *Proceedings of IEEE Pacific-Rim Conference on Multimedia*, 2003.

[61] S. J. Kim, J. Baberjee, W. Kim, and J. F. Garza, "Clustering a Dag for Cad Databases," *IEEE Transactions on Software Engineering*, vol. 14, no. 11, pp. 1684C1699, 1988.

[62] D.-H. Kim and C.-W. Chung, "QCluster: Relevance Feedback Using Adaptive Clustering for Content-based Image Retrieval," in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 2003,

[63] H. Kosch and M. Doller, "Multimedia Database Systems," in *The IASTED International Conference on Databases and Applications*, Innsbruck, Austria, February 2005.

[64] D. Kossmann, "The State of the Art in Distribute Query Processing," *ACM Computing Surveys*, vol. 32, no. 4, pp. 422-469, December 2000.

[65] P. Kumar, S. Ranganath, W. Huang, and K. Sengupta, "Framework for Real-Time Behavior Interpretation from Traffic Video," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 43-53, 2005.

[66] S.-W. Lee, Y.-M. Kim, and S.-W. Choi, "Fast Scene Change Detection using Direct Feature Extraction from MPEG compressed Videos," *IEEE Transaction on Multimedia*, vol. 2, no. 4, pp. 240-254, 2000.

[67] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic Indexing of Soccer Audio-visual Sequences: A Multimodal Approach based on Controlled Markov Chains," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 634-643, 2004.

[68] M. S. Lew, "Next-Generation Web Searches for Visual Content," *IEEE Computer*, vol. 33, pp. 46-53, 2000.

[69] Y. Li, C.-C. J. Kuo, and X. Wan, "Introduction to Content-Based Image Retrieval - Overiew of Key Techniques," edited by V. Castelli and L.D. Bergman, *Image Databases: Search and Retrieval of Digital Imagery*, John Wiley & Sons, Inc., New York, ISBN: 0471321168, pp. 261-284, 2002.

[70] J. Li, J. Z. Wang, and G. Wiederhold, "Simplicity: Semantics-sensitive Integrated Matching for Picture Libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[71] H. C. Lin, L. L. Wang, and S. N. Yang, "Color Image Retrieval Based On Hidden Markov Models, *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 332-339, 1997.

[72] Z. Liu, Y. Wang, and T. Chen, "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 20, no. 1/2, pp. 61-80, 1998.

[73] LSCOM Lexicon Definitions and Annotations Version 1.0, *DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia*, Columbia University ADVENT Technical Report #217-2006-3, March 2006.

[74] G. Lu, Multimedia Database Management Systems. Artech House Publishers, Boston/London, ISBN: 0890063427, 1999.

[75] G. Lu, "Techniques and Data Structures for Efficient Multimedia Retrieval Based on Similarity," *IEEE Transactions on Multimedia*, vol. 4, no. 3, pp. 372-384, 2002.

[76] Y. Lu, H. J. Zhang, W. Liu, and C. Hu, "Joint Semantics and Feature Based Image Retrieval Using Relevance Feedback," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 339-347, 2003.

[77] W.-Y. Ma and H. J. Zhang, “Content-Based Image Indexing and Retrieval,” Handbook of Multimedia Computing, CRC Press, Chapter 13, 1999.

[78] K. Z. Mao, “Feature Subset Selection for Support Vector Machines Through Discriminative Function Pruning Analysis,” *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, vol. 34, no. 1, pp. 60-67, 2004.

[79] J. Mao and A. K. Jain, “A Self-organizing Network for Hyperellipsoidal Clustering,” *IEEE Transactions on Neural Networks*, vol. 7, no. 1, pp. 16-29, 1996.

[80] F. Melgani and L. Bruzzone, “Classification of Hyperspectral Remote Sensing Images with Support Vector Machines,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778-1790, 2004.

[81] M. Mentzelopoulos and A. Psarrou, “Key-Frame Extraction Algorithm Using Entropy Difference,” in *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 39-45, 2004.

[82] MPEG Requirement Group, MPEG-7 Visual Part of Experimentation Model Version 2.0, *Doc. ISO MPEG N2822, MPEG Vancouver Meeting*, 1999.

[83] M. R. Naphade and T. S. Huang, “A Probabilistic Framework for Semantic Indexing and Retrieval in Video,” *IEEE Transactions on Multimedia*, vol. 3, no. 1, March 2001.

[84] A. Natsev, R. Rastogi, K. Shim, “WALRUS: A Similarity Retrieval Algorithm for Image Databases,” *IEEE Trans. on Knowledge and Data Engineering*, vol. 16, No. 3, pp. 301-316, 2004.

[85] V. E. Ogle, “Chabot: Retrieval from a Relational Database of Images,” *Computer*, pp. 40-48, 1995.

[86] M. Ortega, et al., “ Supporting Similarity Queries in MARS,” in *Proceedings of ACM Conferences on Multimedia*, pp. 403-413, 1997.

[87] K. Otsuji and Y. Tonomura, “Projection Detection Filter for Video Cut Detection,” in *Proceedings of ACM Multimedia*, pp. 251-258, 1993.

[88] A. Pentland, R. W. Picard, and A. Sclaroff, “Photobook: Content Based Manipulation of Image Databases,” *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233-254, 1996.

[89] G. Pass, “Comparing Images Using Color Coherence Vectors,” in *Proceedings of ACM International Conference on Multimedia*, pp. 65-73, 1997.

[90] R. J. Povinelli and X. Feng, “A New Temporal Pattern Identification Method for Characterization and Prediction of Complex Time Series Events,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 2, pp. 339-352, 2003.

[91] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[92] L. R. Rabiner and B. H. Huang, “An Introduction to Hidden Markov Models,” *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4-16, 1986.

[93] Y. Rui, A. Gupta, and A. Acero, “Automatically Extracting Highlights for TV Baseball Programs,” in *Proceedings of ACM Multimedia*, pp. 105-115, 2000.

[94] Y, Rui, T. S. Huang, and S. Mehrotra, “Content-based Image Retrieval with Relevance Feedback in MARS,” in *Proceedings of the 1997 International Conference on Image Processing*, pp. 815-818, 1997.

[95] Y, Rui, T. S. Huang, and S. Mehrotra, “Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval,” *IEEE Transactions on Circuit and Video Technology, Special Issue on Segmentation, Description, and Retrieval of Video Content*, vol. 8, no. 5, pp. 644-655, 1998.

[96] Y. Sakurai, M. Yoshikawa, S. Uemura, and H. Kojima, “The A-tree: An Index Structure for High-Dimensional Spaces Using Relative Approximation,” in *Proceedings of the International Conference on Very Large Data Bases*, pp.516-526, 2000.

[97] I. K. Sethi and I. L. Coman, “Mining Association Rules Between Low-Level Image Features and High-Level Concepts,” in *Proceedings of SPIE Data Mining and Knowledge Discovery*, vol. 3, pp. 279-290, 2001.

[98] M.-L. Shyu, S.-C. Chen, M. Chen, and C. Zhang, “Affinity Relation Discovery in Image Database Clustering and Content-Based Retrieval,” in *Proceedings of ACM International Conference on Multimedia*, pp. 372-375, 2004.

[99] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang, and K. Sarinnapakorn, “Image Database Retrieval Utilizing Affinity Relationship,” in *Proceedings of the First ACM International Workshop on Multimedia Databases*, pp. 78-85, 2003.

[100] M.-L. Shyu, S.-C. Chen, M. Chen, C. Zhang, and C.-M. Shu, “Probabilistic Semantic Network-based Image Retrieval Using MMM and Relevance Feedback,” *Journal of Multimedia Tools and Applications*, vol. 30, no. 2, pp. 131-147, 2006.

[101] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, “Mining User Access Behavior on the WWW,” in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1717C1722, 2001.

[102] M.-L. Shyu, S.-C. Chen, and C. Haruechaiyasak, C.-M. Shu, and S.-T. Li, “Disjoint Web Document Clustering and Management in Electronic Commerce,” *Proceedings of the Seventh International Conference on Distributed Multimedia Systems (DMS2001)*, pp. 494-497, 2001.

[103] M.-L. Shyu, S.-C. Chen, and R.L. Kashyap, "Database Clustering and Data Warehousing," in *Proceedings of the 1998 ICS Workshop on Software Engineering and Database Systems*, pp. 30-37, 1998.

[104] M.-L. Shyu, S.-C. Chen, and R.L. Kashyap, "A Probabilistic-Based Mechanism for Video Database Management Systems," *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2000)*, pp. 467-470, 2000.

[105] M.-L. Shyu, S.-C. Chen, and R.L. Kashyap, "Organizing a Network of Databases Using Probabilistic Reasoning," *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1990-1995, 2000.

[106] M.-L. Shyu, S.-C. Chen, and S. H. Rubin, "Stochastic Clustering for Organizing Distributed Information Source," *IEEE Transactions on Systems, Man and Cybernetics: Part B*, vol. 34, no. 5, pp. 2035-2047, 2004.

[107] J. R. Smith and S. F. Chang, "Automated Image Retrieval Using Color and Texture," *Technical Report CU/CTR 408-95-14*, Columbia University, July 1995.

[108] J. R. Smith and S. F. Chang, "VisualSeek: A Fully Automated Content-Based Query System," *Proceedings of ACM Multimedia*, pp. 87-98, 1996.

[109] R. O. Stehling, M. A. Nascimento, and A. X. Falcao, "On Shapes of Colors for Content-Based Image Retrieval," in *Proceedings of ACM International Workshop on Multimedia Information Retrieval*, pp. 171-174, 2000.

[110] G. Stumme, R. Taouil, et al., "Computing Iceberg Concept Lattices with Titanic," *Data and Knowledge Engineering*, vol. 42, no. 2, pp. 189-222, 2002.

[111] C.-W. Su, H.-Y. M. Liao and K.-C. Fan, "A Motion-Flow-Based Fast Video Retrieval System," in *Proceedings of 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 105-112, 2005.

[112] X.Sun and M.Kankanhalli, "Video Summarization Using R-Sequences," *Real-time Imaging*, vol. 6, no. 6, pp. 449-459, 2000.

[113] H. Sun, J.-H. Lim, Q. Tian, and M. S. Kankanhalli, "Semantic Labeling of Soccer Video," in *Proceedings of IEEE Pacific-Rim Conference on Multimedia*, pp. 1787-1791, 2003.

[114] D. Swanberg, C. F. Shu, and R. Jain, "Knowledge Guided Parsing in Video Database," in *Proceedings of SPIE'93, Storage and Retrieval for Image and video Databases*, vol. 1908, pp. 13-24, 1993.

[115] P.-N. Tan, et al., Introduction to Data Mining, Addison Wesley, ISBN: 0-321-32136-7.

[116] Y. Tan and J. Wang, "A Support Vector Machine with a Hybrid Kernel and Minimal Vapnik-Chervonenkis Dimension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 385-395, 2004.

[117] D. Tjondronegoro, Y.-P. Chen, B. Pham, "Content-based Video Indexing for Sports Analysis," *Proceedings of ACM Multimedia*, pp. 1035-1036, 2005.

[118] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," in *proceedings of ACM International Conference on Multimedia*, pp. 107-118, 2001.

[119] X. Tong, L. Duan, et al., "A Mid-level Visual Concept Generation Framework for Sports Analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 646-649, 2005.

[120] V. Tovinkere and R. J. Qian, "Detecting Semantic Events in Soccer Games: Towards A Complete Solution," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1040-1043, 2001.

[121] R. Vilalta and S. Ma, "Predicting Rare Events in Temporal Domains," in *Proceedings of IEEE International Conference on Data Mining*, pp. 474-481, 2002.

[122] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/

[123] K. Wan, X. Yan, X. Yu, and C. Xu, "Real-time Goal-mouth Detection in MPEG Soccer Video," in *Proceedings the 11th ACM International Conference on Multimedia*, pp. 311-314, 2003.

[124] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-Sensitive Integrated Maching for Picture Libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947-963, 2001.

[125] Y. Wang, Z. Liu, and J. Huang, "Multimedia Content Analysis Using Both Audio and Visual Clues,"*Signal Processing Magazine*, vol. 17, pp. 12-36, 2000.

[126] J. Wang, C. Xu, E. Chng, K. Wah, and Q. Tian, "Automatic Replay Generation for Soccer Video Broadcasting," in *Proceedings of the 12th ACM International Conference on Multimedia*, pp. 311-314, 2004.

[127] A.K. Wasfi and G. Arif, "An Approach for Video Meta-Data Modeling and Query Processing," in *Proceedings of ACM Multimedia*, pp. 215-224, 1999.

[128] R. Weber, H.-J. Schek, and S. Blott, "A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces," in *Proceedings of the International Conference on Very Large Databases (VLDB)*, pp. 194-205, 1998.

[129] Q. Wei, H. Zhang, and Y. Zhong, "Robust Approach to Video Segmentation Using Compressed Data," in *Proceedings of the Conference on Storage and Retrieval for Image and Video Database*, vol. 3022, pp. 448-456, 1997.

[130] G. Wiederhold, "Mediators in the Architecture of Future Information Systems," *IEEE Computers*, pp. 38-49, 1992.

[131] C. Wu, et al., "Event Recognition by Semantic Inference for Sports Video," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 805-808, 2002.

[132] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Unsupervised Discovery of Multilevel Statistical Video Structures using Hierarchical Hidden Markov Models," in *Proceedings of IEEE International Conference on Multimedia and Expo*, vol. 3, pp. 29-32, 2003.

[133] M. Xu, et al., "Creating Audio Keywords for Event Detection in Soccer Video," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 281-284, 2003.

[134] P. Xu, L. Xie, S.-F. Chang, et al., "Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 928-931, 2001.

[135] K. Yanai, "Generic Image Classification Using Visual Knowledge on the Web," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, pp. 167-176, 2003.

[136] Z. Yang, X. Wang, and C.-C. J. Kuo, "Interactive Image Retrieval: Concept, Procedure and Tools," in *Proceedings of the IEEE* 32^{th} *Asilomar Conference*, pp. 261-265, 1998.

[137] Q. Ye, Q. Huang, W. Gao, and S. Jiang, "Exciting Event Detection in Broadcast Soccer Video with Mid-level Description and Incremental Learning," in *Proceedings of ACM Multimedia*, pp. 455-458, 2005.

[138] H. H. Yu and W. Wolf, "Multiresolution Video Segmentation Using Wavelet Transformation," in *Proceedings of the Conference on Storage and Retrieval for Image and video Database*, vol. 3312, pp. 176-187, 1998.

[139] X. Yu, C. Xu, et al. "Trajectory-based Ball Detection and Tracking with Applications to Semantic Analysis of Broadcast Soccer Video," in *Proceedings of the 11th ACM International Conference on Multimedia*, pp. 11-20, 2003.

[140] D. Q. Zhang and S.-F. Chang, "Event Detection in Baseball Video using Superimposed Caption Recognition," in *Proceedings of the 10th ACM International Conference on Multimedia*, pp. 315-318, 2002.

[141] C. Zhang, S. C. Chen, and M. L. Shyu, "Multiple Object Retrieval for Image Databases Using Multiple Instance Learning and Relevance Feedback," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 775-778, 2004.

[142] H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic Partitioning of Full-Motion Video," *Multimedia Systems*, vol. 1, no. 1, pp. 10-28, 1993.

[143] D. S. Zhang and G. Lu, "Generic Fourier Descriptors for Shape-Based Image Retrieval," in *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 425-428, August 2002.

[144] C. Zhang, S.-C. Chen, M.-L. Shyu, "PixSO: A System for Video Shot Detection," in *Proceedings of the Fourth IEEE Pacific-Rim Conference on Multimedia*, pp. 1-5, 2003.

[145] D. Zhong and S.-F. Chang, "Content Based Video Indexing Techniques," ADVENT Research Report.

[146] X. S. Zhou, Y. Rui, and T. Huang, "Water-Filling: a Novel Way for Image Structural Feature Extraction," in *Proceedings of IEEE International Conference on Image Processing*, vol. 2, pp. 570-574, 1999.

[147] W. Zhou, A. Vellaikal, and C.-C. J. Kuo, "Rule-Based Video Classification System for Basketball Video Indexing," in *Proceedings of ACM International Conference on Multimedia*, pp. 213-216, 2000.

[148] X. Zhu, et al., "Video Data Mining: Semantic Indexing and Event Detection from the Association Perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 5, pp. 665-677, 2005.

[149] R. Zabih, J. Miller, and K. Mai, "A Feature-based Algorithm for Detecting and Classifying Scene Breaks," in *Proceedings of ACM Multimedia*, pp. 189-200, 1995.

VITA

MIN CHEN

July 20, 1976 Born, Fenghua, Zhejiang, P. R. China

1997 B.E., Electrical Engineering
Zhejiang University, P. R. China

1997 – 2001 Motorola Cellular Equipment Co., Ltd.
Zhejiang, P. R. China

2004 M.E., Computer Science
Florida International University, Miami, Florida

2004 – 2007 Doctoral candidate, Computer Science
Florida International University, Miami, Florida

PUBLICATIONS AND PRESENTATIONS

Chen, S.-C., Zhang, K., and Chen, M. (2003). "A Real-Time 3D Animation Environment for Storm Surge," in *Proc. of the IEEE Intl. Conf. on Multimedia & Expo*, vol. I, pp. 705-708.

Chen, S.-C., Shyu, M.-L., Zhang, C., Luo, L., and Chen, M. (2003). "Detection of Soccer Goal Shots Using Joint Multimedia Features and Classification Rules," in *Proc. of the 4th Intl. Workshop on Multimedia Data Mining*, pp. 36-44.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2003). "Image Database Retrieval Utilizing Affinity Relationships," in *Proc. of the 1st ACM Intl. Workshop on Multimedia Databases*, pp. 78-85.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2003). "MMM: A Stochastic Mechanism for Image Database Queries," in *Proc. of the IEEE 5th Intl. Symposium on Multimedia Software Engineering*, pp. 188-195.

Chen, S.-C., Shyu, M.-L., Chen, M., et al. (2004). "A Decision Tree-based Multimodal Data Mining Framework for Soccer Goal Detection," in *Proc. of IEEE Intl. Conf. on*

Multimedia and Expo, vol. 1, pp. 265-268.

Chen, S.-C., Hamid, S., Gulati, S., Zhao, N., Chen, M., et al. (2004). “A Reliable Web-based System for Hurricane Analysis and Simulation,” in *Proc. of IEEE Intl. Conf. on Systems, Man and Cybernetics*, pp. 5215-5220.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2004). “Affinity Relation Discovery in Image Database Clustering and Content-based Retrieval,” in *Proc. of ACM Multimedia 2004 Conference*, pp. 372-375.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2004). “A Unified Framework for Image Database Clustering and Content-based Retrieval,” in *Proc. of ACM Intl. Workshop on Multimedia Databases*, pp. 19-27.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2004). “Affinity-Based Similarity Measure for Web Document Clustering,” in *Proc. of IEEE Intl. Conf. on Information Reuse and Integration*, pp. 247-252.

Zhang, C., Chen, X., Chen, M., et al. (2005). “A Multiple Instance Learning Approach for Content Based Image Retrieval Using One-Class Support Vector Machine,” in *Proc. of IEEE Intl. Conf. on Multimedia and Expo*, pp. 1142-1145.

Chen, X., Zhang, C., Chen, S.-C., and Chen, M. (2005). “A Latent Semantic Indexing Based Method for Solving Multiple Instance Learning Problem in Region-based Image Retrieval,” in *Proc. of IEEE Intl. Symposium on Multimedia*, pp. 37-44.

Wickramaratna, K., Chen, M., Chen, S.-C., and Shyu, M.-L. (2005). “Neural Network Based Framework for Goal Event Detection in Soccer Videos,” in *Proc. of IEEE Intl. Symposium of Multimedia*, pp. 21-28.

Chen, M. and Chen, S.-C. (2006). “MMIR: An Advanced Content-based Image Retrieval System using a Hierarchical Learning Framework ,” Edited by Zhang, D. and Tsai, J. *Advances in Machine Learning Application in Software Engineering*, Idea Group Publishing, ISBN: 1-59140-941-1.

Chen, M., et al. (2006). “Semantic Event Detection via Temporal Analysis and Multimodal Data Mining,” *IEEE Signal Processing Magazine*, Special Issue on Semantic Retrieval of Multimedia, vol. 23, no. 2, pp. 38-46.

Chen, S.-C., Shyu, M.-L., Zhang, C. and Chen, M. (2006). “A Multimodal Data Mining Framework for Soccer Goal Detection Based on Decision Tree Logic,” *Intl. Journal of Computer Applications in Technology*, vol. 27, no. 4, pp. 312-323.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2006). “Probabilistic Semantic Network-based Image Retrieval Using MMM and Relevance Feedback,” *Multimedia Tools and*

Applications, vol. 30, no. 2, pp. 131-147.

Chatterjee, K., Saleem, K., Zhao, N., Chen, M., et al. (2006). “Modeling Methodology for Component Reuse and System Integration for Hurricane Loss Projection Application,” in *Proc. of IEEE Intl. Conf. on Information Reuse and Integration*, pp. 57-62.

Chen, S.-C., Chen, M., et al. (2006). “Exciting Event Detection using Multi-level Multimodal Descriptors and Data Classification,” in *Proc. of IEEE Intl. Symposium on Multimedia*, pp. 193-200.

Shyu, M.-L., Chen, S.-C., Chen, M., et al. (2007). “Capturing High-Level Image Concepts via Affinity Relationships in Image Database Retrieval,” *Multimedia Tools and Applications*, vol. 32, no. 1, pp. 73-92.

Chen, M., et al. (2007). “Video Event Mining via Multimodal Content Analysis and Classification,” Edited by Petrushin, V. A. and Khan, L. *Multimedia Data Mining and Knowledge Discovery*, Springer Verlag, ISBN: 978-1-84628-436-6.

Chen, M., et al. (accepted). “Hierarchical Temporal Association Mining for Video Event Detection in Video Databases,” accepted for publication, *IEEE Intl. Workshop on Multimedia Databases and Data Management*, in conjunction with IEEE International Conference on Data Engineering, Istanbul, Turkey.

Zhao, N., Chen, M., et al. (accepted). “User Adaptive Video Retrieval on Mobile Devices,” accepted for publication, Edited by Yang, L. T., Waluyo, A. B., Ma, J., Tan, L. and Srinivasan, B. *Mobile Intelligence: When Computational Intelligence Meets Mobile Paradigm*, John Wiley & Sons Inc.

Chen, X., Zhang, C., Chen, S.-C. and Chen, M. (accepted). “LMS - A Long Term Knowledge-Based Multimedia Retrieval System for Region-Based Image Databases,” accepted for publication, *Intl. Journal of Applied Systemic Studies.*