COMBINING IMAGE FEATURES FOR SEMANTIC DESCRIPTIONS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

MEDENİ SOYSAL

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2003

Approval of the Graduate School of Natural and Applied Sciences

_____

Prof. Dr. Canan Özgen

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Prof. Dr. Mübeccel Demirekler

Head Of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. A.Aydın Alatan

Supervisor

Examining Committee Members

Prof. Dr. Mübeccel Demirekler	_____
Prof. Dr. Uğur Halıcı	_____
Assoc. Prof. Dr. A. Aydın Alatan	_____
Assoc. Prof. Dr. Gözde Bozdağı Akar	_____
Ersin Esen, M.S.	_____

ABSTRACT

COMBINING IMAGE FEATURES FOR SEMANTIC DESCRIPTIONS

Soysal, Medeni

MSc., Department of Electrical and Electronics Engineering

Supervisor: Associate Professor A.Aydın Alatan

September 2003, 85 pages

Digital multimedia content production and the amount of content present all over the world have exploded in the recent years. The consequences of this fact can be observed everywhere in many different forms, to exemplify, huge digital video archives of broadcasting companies, commercial image archives, virtual museums, etc. In order for these sources to be useful and accessible, this technological advance must be accompanied by the effective techniques of indexing and retrieval. The most effective way of indexing is the one providing a basis for retrieval in terms of semantic concepts, upon which ordinary users of multimedia databases base their queries. On the other hand, semantic classification of images using low-level features is a challenging problem. Combining experts with different classifier structures, trained by MPEG-7

low-level color and texture descriptors, is examined as a solution alternative. For combining different classifiers and features, advanced decision mechanisms are proposed, which utilize basic expert combination strategies in different settings. Each of these decision mechanisms, namely Single Feature Combination (SFC), Multiple Feature Direct Combination (MFDC), and Multiple Feature Cascaded Combination (MFCC) enjoy significant classification performance improvements over single experts. Simulations are conducted on eight different visual semantic classes, resulting in accuracy improvements between 3.5-6.5%, when they are compared with the best performance of single expert systems.

Keywords: content-based indexing, MPEG-7, combining classifiers

ÖZ

ANLAMSAL TANIMLAMALAR İÇİN GÖRÜNTÜ ÖZNİTELİKLERİ BİRLEŞTİRME

Soysal, Medeni

Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi : Doç. Dr. A. Aydın Alatan

Eylül 2003, 85 sayfa

Sayısal çoğulortam içeriği üretimi ve dolayısıyla tüm dünyada varolan içerik miktarı geçtiğimiz yıllarda büyük bir artış göstermiştir. Bu artışın sonuçlarına değişik şekillerde de olsa her yerde rastlanabilmektedir. Örnek vermek gerekirse, belli başlı yayın kuruluşlarının dev sayısal video arşivleri, ticari amaçlı imge arşivleri ve sanal müzeleri bunlar arasında sayabiliriz. Bu kaynakların yararlı ve erişilebilir olması için, bu teknolojik ilerlemenin etkili dizinleme ve erişim teknikleriyle desteklenmesi gerekmektedir. En etkili dizinleme, çoğulortam veritabanlarının sıradan kullanıcılarının sorgularını dayandırdıkları anlamsal kavramları temel alan erişime olanak sağlayan dizinlemedir. Öte yandan, imgelerin düşük seviyeli tanımlayıcılar yoluyla anlamsal sınıflara ayrılması zor bir problemdir. Değişik sınıflandırıcı yapılarına sahip, MPEG-7 düşük seviye

renk ve doku tanımlayıcıları ile eğitilmiş uzmanların birleştirilmesi bir çözüm alternatifi olarak incelenmiştir. Değişik sınıflandırıcı yapılarını ve imge özelliklerini biraraya getirebilmek için temel uzman birleştirme metodlarını farklı şekillerde kullanan gelişmiş karar mekanizmaları önerilmiştir. Bu karar mekanizmalarının herbiri sınıflandırma performansı konusunda tekil uzmanların sağladığından daha başarılı sonuçlar elde etmişlerdir. Deneyler sekiz ayrı görsel anlamsal sınıf üzerinde yapılmış ve birleşik uzmanlar tekil uzmanların en başarılısından %3.5-6.5 arasında daha iyi sonuç vermişlerdir.

Anahtar Kelimeler: içerik tabanlı dizinleme, MPEG-7, sınıflandırıcı birleştirme.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. A. Aydın Alatan for sharing his experience and valuable ideas with me. His guidance and insight has always enlightened my way during the last two years. My colleague Ersin Esen has also offered his valuable help to me generously during the research period of this thesis, and will be remembered with appreciation. I also want to thank my "brothers in arms", Özgür Deniz Önür and Yağız Yaşaroğlu, with whom every challenge has been a pleasure. The rest of the Enkare members who played an active role in this thesis, Ekin Dino, Çağlar Karasu, and especially Oktay Onur Kuzucu will also be remembered with special words. Without them the preparation period of this thesis might have been shorter, but surely it would have been far more dull.

Last, but not the least, I am grateful to my family for all their support and sacrifices. My mother, Ayşe, for her sincere encouragement and help in every important task of my life, and my father, Mehmet Medeni, for his adorable character and estimable life that he left behind as the ideals that I will never yield to follow.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In the recent years, technological advances have made multimedia content production easier than ever before. Digital cameras and many other personal recording devices that enable content production are pervasively utilized today. By Internet, it also became possible to exchange produced content throughout the world, multiplying the amount of data accessible.

Until recently, price of storage devices was a limiting factor for the amount of content that is possessed and utilized. However, this fact also changed by the quick advances in the storage technology. As an illustration, consider the amount of video (MPEG-1) that can be stored in a hard disk that's worth $100. Today, approximately 100 hours of video can be stored with a cost of $100, while two years ago the amount that could be stored was only 5 hours.

As a result of these great strides in content production, storage and exchange, large collections of digital multimedia data are used in various areas today [1]. These areas include planning, government, and military intelligence that use satellite imagery, as well as commercial photo libraries and digital museums that have extremely large and very diverse image collections.

In addition to these specific areas, there is a huge amount of disorganized content that is produced and shared by common users (e.g. home video).

Unfortunately, what would seem like a great technological jump and a pleasant dream can easily turn into a chaos, if no sophisticated tools for managing the explosion in available content are developed. It should be remembered that information has value only if it can be reached and consumed. These solid proofs have led researchers to the common consensus that indexing and management is compulsory for digital multimedia data to be valuable in the long term.

1.1 Scope of the Thesis

The main idea behind this thesis research is reaching high-level semantic descriptions of still images by utilizing low-level features.For this purpose, combining the results of different classifier structures trained by various low-level features is proposed.

Proposed method is tested on eight different visual semantic classes, which are football, indoor, crowd, sunset-sunrise, sky, forest, sea, and cityscape.

1.2 Content Based Information Retrieval

The ease in production and storage of multimedia data increases the speed that the amount of the data grows and hence makes the problem of managing more complex. In fact, the complexity lies in building appropriate representations of the data, since indexing is not possible without such a representation. This fact can be best explained by the famous example [1]: Imagine a large secondhand book shop whose books are sorted by the color of the dust-jacket. No matter how large the

collection would be, this shop could attract very few customers and these would probably be the most desperate ones.

As a primary fact, indexing and retrieval should be based on the content. The usual strategies of people, who own commercial image and video databases, are based upon manual annotation. This means annotating each picture or video by hand and it is increasingly unacceptable for two reasons. First of all, it is a costly process and if one considers the growing amount of the content, it is obvious that the volume of work involved will be tremendous. Secondly, preparing an objective and domain-independent description by using an inherently subjective method that involves human annotators is not possible.

Automatic annotation is an alternative to manual indexing. It involves analysis of data by automatically extracting features that will be useful for searching and discovering content. Representation of data by means of automatically extracted low-level descriptors has three main advantages over manual annotations: (1) they will be automatically extracted without consuming human work power, (2) they can be more objective and domain-independent and (3) they can be native to the multimedia data, meaning that they do not involve textual descriptions, but use features, such as color, texture and shape.

The next step in automatic annotation is reaching the semantic descriptions using the extracted features. At this point, researchers are faced with the most challenging task: Bridging the gap between low-level descriptors that machines can process and the high-level descriptions that users can understand and use. This thesis aims to shorten this gap.

1.3 A Brief Review of MPEG-7 Standard

The problem mentioned and the technological situation were also recognized by International Standards Organization Moving Pictures Expert Group (ISO MPEG). Considering the growing interest towards the management of multimedia data in the recent years, ISO MPEG organization established a new standard, MPEG-7, for describing various types of multimedia data [26].

Similar to the other MPEG standards, namely MPEG-1, MPEG-2 and MPEG-4, MPEG-7 defines a standard representation of multimedia information with a set of well-defined requirements. However, MPEG-7 substantially differs from all other MPEG standards due to its target. While all the others represent the content itself, MPEG-7 represents information about the content. Therefore, it is not a coding standard, but a multimedia content description interface. Moreover, it should also be noted that MPEG-7, does not standardize how the information is to be extracted or consumed, whereas it standardizes which information is to be extracted and utilized.

MPEG-7 standard comes up with many features, supporting both manual and automatic annotation alternatives. In its context, although many detailed media descriptions for manual annotation exist, automatic annotation is strongly encouraged by many audiovisual low-level descriptors based on native properties of the multimedia content. (i.e. color, texture, shape, melody, etc.) The most important feature which makes MPEG-7 different and probably superior than other digital management methods is that, it addresses the interoperability issues and therefore provides a standard way to do multimedia content description and allows the exchange of content and its descriptions across different systems.

1.4 Outline

The main focus in this research is about Content Based Image Retrieval (CBIR) using low-level visual features. Tools, methods, approaches and systems from the related literature about this subject is investigated in detail in the next chapter. A realization of a system, whose main aim is to integrate methods and tools used in many systems so far, with some newly emerged ones, is presented in Chapter 3. This system combines the outputs of many different complex mechanisms to reach high-level concepts that ordinary users may seek in their queries. In Chapter 4, simulation results of the system proposed are presented. Simulation results are discussed in detail and some concluding remarks are made in the final chapter. Moreover, an outline of the future research will be given, along with the ideas emerged during this research.

CHAPTER 2

STATE OF THE ART IN CONTENT BASED IMAGE RETRIEVAL

State-of-the-art Content Based Image Retrieval Systems (CBIR) are based on techniques and approaches which are results of long and intensive research conducted on this subject. In this chapter, basic components of these systems are analyzed as a detailed summary of the literature. These basic components can be listed as low-level features, the distance measures used for discriminating features, and similarity and classification based retrieval methods. In the last section, a review of current content based retrieval systems will also be presented [2]. This review describes the most popular commercial and academic CBIR systems as well as some systems enjoying interesting and novel features. This list can be considered to contain the key players in the area.

2.1 Low-level Features

Low-level feature can be defined as a description of a multimedia signal, which has undergone minimum processing after being captured by sensors. When the multimedia signal under consideration is a still image, applicable low-level features are color, texture and shape. Following sub-sections are devoted to the

literature about these three low-level features which are utilized for describing still images according to their visual content.

2.1.1 Color Descriptors

Color being an important visual attribute for human vision is used extensively in image retrieval. It can be defined in many different spaces such as RGB, YCbCr, HSV, HMMD, etc. Many different techniques for comparing images in terms of color similarity have been described in the literature [2].

The first and the most common technique used in indexing of images according to their color content is the Color Histogram [3]. In this technique, colors in the image are mapped into a discrete color space containing a predefined number of colors and the number of points mapped to each point in this new space is calculated. It is used in many image retrieval systems [3,4], due to its success in characterizing the global color content of the image.

Experiments with the color histogram technique have proved that this technique, though being very useful, is vulnerable against the quantization parameter of the histograms. Considering this property, a variation of this technique, namely Cumulated Color Histogram, is proposed [3]. Cumulated Color Histogram, as it can be inferred from its name, calculates the number of image pixels cumulatively and proved more robust than the first one.

Another important approach is Color Moments technique [5]. This technique proposes the use of only the dominant features of the image color distribution with a small representative set of color vectors that capture the color properties of the image. In order to provide a faster search tool, Color Sets are

proposed [2]. This approach identifies the regions within the image containing colors from a predefined set. Image is represented by binary vectors corresponding to image regions.

Above methods provide a basis for the research in color-based querying. However, they have limited capability of similarity matching due to the lack of spatial information incorporation. There are also techniques which are based on both spatial relations and color feature. The easiest way to include spatial information is Sub-block Histogram [6]. In this method, image is divided into a predefined number of sub-blocks and color histogram is computed for each of these. Then, similarity search is done by calculating the histogram difference between the corresponding blocks.

Region-based color querying is another technique, which enables selecting a region inside an image for using its features in similarity search [2,7,8]. This technique has the drawback of requiring a preprocessing operation for segmenting the search image into regions. On the other hand, Spatial Chromatic Histograms incorporate the information provided by color histogram with the information about the location of pixels of similar color and their arrangement within the image [2]. Finally, Color Correlograms are also proposed as color features. These features can be used to describe the global distribution of local correlations, since they include the spatial correlation of colors.

All of the techniques above are developed in order to model the perceptual similarity between the query and target images. Although they provide powerful tools for color-based matching, images with quite different appearance may still be considered as “similar” because of similar color compositions. Modeling the

human visual system by using algorithms that computers can perform is still a very popular topic.

2.1.2 Texture Descriptors

Texture is the repetition of a basic pattern over a given area. Its repetition rate, as well as the shape of the pattern, defines texture features. In other words, texture of a visual item characterizes the interrelationship between its adjacent pixels. Texture similarity is a quite useful property when distinguishing images that have similar color content (e.g. sky-sea, leaves-grass, etc.). There are many techniques to reveal the texture features that have significance for human eye. All of these techniques work on the grayscale representations of the image pixels.

Since texture is a property recognized by the human eye, techniques for modeling the human perception of texture are also developed [2]. Psychological studies have proved that coarseness, contrast, directionality, regularity and roughness are the properties mainly used by humans to define texture [3]. Considering this, many computational approximations were developed for defining these important visual texture properties [2,9].

A well known and inspiring technique for defining texture is Co-occurrence Matrix [2]. This technique computes a two dimensional histogram of the dependencies of neighboring grayscale values. Frequency transform based methods are also very popular in this area. A number of methods using the results of 2D DFT and 2D DCT applied on images are the first examples of the research on this subject [10].

After wavelet transform [3] became popular among researchers, it is also integrated into texture analysis research. While in one technique, mean and variance extracted from the wavelet subbands are used for texture representation [3], wavelet transform is combined with the co-occurrence matrix in another [2]. Wavelet-based histogram technique being a variation of the first, is also proposed to use a coefficient histogram, utilizing channels as indexes [11]. Gabor Wavelet Transform is also used in this subject, since it very closely models the human vision of texture. An example, based on this transform, computes coefficients of a codebook of important texture patterns and then retrieves images, according to their similarity [2].

Edge Direction Histogram is a technique that has recently become widespread [12]. In this technique, edges with predefined basic orientations are counted inside images. Afterwards, in order to compute similarity, histograms consisted of the number of edges are used.

2.1.3 Shape Descriptors

Shape is an important feature to identify the similarity in a binary image material. In order to obtain a more complex semantic representation, images should be analyzed based on objects. This can only be achieved via shape descriptors. Since shape properties can only be extracted from binary images, input image should undergo a segmentation process beforehand.

During the long studies on defining metrics for shape, many different descriptors are proposed. Some of these descriptors are area, center of mass, circularity, moments, length, irregularity, complexity, aspect ratio, right

angleness, sharpness and directedness [2,9]. These descriptors can be considered as parameters that are used in shape recognition by the human vision system.

Descriptors for shape representation in images are divided into two main categories which are region-based and contour-based. The most successful representation in the category of region-based descriptors is reported as the Moment Invariants. Seven such moments, which are known as *Hu Moments* [2], are proved to be successful, especially when invariance to transformation is considered. On the other hand, Curvature Scale Space Representation is the most popular among the contour-based descriptors [2,13]. This representation locates the strongest peaks on the contour and computes their strength in order to reveal the characteristics of the contour.

2.2 Distance Measures

For CBIR purposes, two multimedia content should be compared in terms of their low-level features. This situation requires a corresponding distance measure for each low-level feature. Low-level image features are usually represented by vectors of various dimensions. Therefore, images can be compared by evaluating the distance of these representative n-dimensional vectors between each other. There are many different distance metrics that are native to a specific descriptor, in addition to the generic ones.

Absolute and Euclidian Distances are simply the well known distance metrics, used in many different areas [2]. They are also known as L1 and L2-norms, respectively. Mahalanobis Distance is also a well known distance metric in statistics [2]. It is used in modeling multivariate distributions and based on

covariance values of the vectors. Baddeley Distance is one of the specific measures that are mentioned above [14]. They are based on distance transforms and known as a strong similarity measure applied specifically to color vectors.

For color histograms, many specific distance measures are also defined. One of them is the Quadratic Distance [9] for which a similarity matrix that accounts for the perceptual similarity of two histogram bins is used. Another common distance measure for color histograms is known as Histogram Intersection, and also known to be a simple method for fast applications [2]. Lastly, revisiting L1 and L2 norms, according to the experiments, when applied to histograms, L1-norm has higher discrimination power than L2-norm.

Shape Descriptors also have a specific distance metric which is called Hausdorff Distance [2]. Hausdorff Distance is simply defined as the maximum distance of a set of features to the nearest point in the other set of features. It is applied to shape similarity for comparing two different sets of areas characterizing the shapes that they belong to. It assigns the highest dissimilarity as the distance between them.

Summing up, one can say that the performance of retrieval mechanisms are closely related to the success of the metric used for matching, in modeling human judgments of similarity.

2.3 Similarity-based Retrieval

The basic idea of similarity-based retrieval is to extract characteristic features from the query image and then compare them against the features of the images in

a particular database to find the most similar images. The utilized features depend on the specific application and must be selected before the query.

This method has many application areas, especially in art and fashion. As an illustration, imagine yourself as a fashion designer who is in need of fabric images with a particular mixture of colors, or a documentary director who requires red Japanese fish photos.

Although, similarity based retrieval has applications in many areas, it is far from solving the CBIR problem completely. This arises from the fact that many users are not experts having past experience about low-level image properties, such as color, texture and shape. Without having expert knowledge about the low-level features, it is not possible to select the feature which best represents the concept conveyed by the image. In addition, the requirement to provide a query image that represents the concept in mind can be quite cumbersome in many cases. As a conclusion, this can be considered as a step towards the solution, but not the final solution itself.

2.4 Classification-based Retrieval

Ordinary users of image databases all over the world do not have specific knowledge about the characteristic features of images that they are interested in. However, all of them have the ability to name the semantic concept that they are looking for. The queried concept ranges from generic to specific. For instance, “Find me indoor images” defines a quite generic query, whereas “Find me photos of Tayyip Erdoğan while he was falling down from a horse” defines a very specific one. As seen in these examples, classification-based retrieval addresses

the much harder problem of bridging the wide gap between semantics that users can define and the low-level features that can be automatically extracted by the computers.

Unfortunately, in an unconstrained environment this is still an unsolved problem. The effectiveness of the solutions to the problem depends highly on the application domain. In the systems that semantic queries are available, a semantic concept should first be defined by a sufficient number of images which contain this concept. After the system is trained by using the provided data, it can classify new test data automatically. However, as it is previously mentioned the success of this kind of system is highly dependent on the ability of the training data to make the system capturing the concept.

Another important fact about semantic queries is that most of the time they are based on more complex concepts than can be inferred from the whole image. This forces the systems to work on objects or regions rather than the whole image. In order to fulfill this kind of requests, object regions inside images should be initially segmented and defined in terms of low-level features separately. However, another unsolved problem, image segmentation, hinders the success at this point.

2.5 Man-in-the-loop

Today's technology in many cases does not let the user to stay completely out of the decision phases and still reach the semantics successfully. Following this idea, many of the approaches are modified to include some feedback from the user.

There are many methods to utilize the user response to a given query. One of the most common approaches is to ask the user to rank the results that are returned by the system to a given query [26]. Another approach is to get some positive and negative feedbacks from the user like “Select the least relevant result” or “Select the most relevant result”. All of these approaches are investigated under the topic *relevance feedback*, and use the response from the user to return refined results back to the user.

In the systems involving segmentation and a query image, user is given the chance to visually specify the target precisely. This is achieved via segmenting the image provided by the user and then asking the user to select the region that is relevant to the concept searched. In addition, there are systems in the literature that give user the chance to specify the importance of low-level features in the matching process.

There are many variations of these approaches for adding the user to the loop. However, one should keep in mind the fact that although user interaction can be very useful, if exceeds a limit, may degrade the value of the system and also veil most of its success.

2.6 Systems

During the last decade many content based indexing and retrieval systems have been developed. Their inherent technologies greatly vary and these systems are currently in use for both general and domain specific applications. None of the technologies or systems that are in use has become pervasive among all others.

In this section, some of the most popular CBIR systems in the literature are introduced. These systems are selected in order to represent different approaches based on input interface, query techniques and indexing features.

One of the first, and probably the most pervasive of CBIR systems is IBM's QBIC (Query By Image Content) [9,15]. It is a commercial system which is a module of the IBM DB2 Database System that is used by museums for their online galleries. It has support for queries based on given example images, selected color and texture patterns and annotation text.

Netra system developed at University of California, is a more experimental image retrieval system [2]. It supports region-based color,texture and shape features for indexing and retrieval. Images are segmented at the input stage and features to be used for indexing and retrieval are extracted from the regions found.

VisualSEEk is the successor of the WebSEEk system that is developed by Columbia University [8]. This system also segments the images at the input stage. For indexing, color sets are used.

Virage Image Engine is another commercial product developed by Virage Organization [16]. It is an "Open Framework" platform which provides developers the necessary tools for extracting image features.

SCHEMA, which is a network of excellence under the EC 5th Framework Information Society Technologies Programme, has also developed a content based image retrieval system named ISTORAMA [17,18]. This system is still under development and at this point provides a web-based interface to the users for testing. This system both supports region-based color, texture and shape queries

enabling the user to adjust the weight of each feature used, and also has a category based simplistic interface.

Caliph&Emir, developed by Know-Center, supports complex manual annotations, as well as queries based on color and texture [19]. It is one of the first systems that utilizes MPEG-7 color and texture features for automatic indexing.

Another tool released by IBM is VideoAnn [20]. This tool has a different interface enabling the user to create lexicons first, and then use these to annotate inputs with these merely checking and unchecking lexicon boxes. It also supports region annotation with the interface mentioned above.

Tecmath developed MediaArchive, which is a powerful archiving tool allowing large amounts of media files to be stored reliably [21]. It is an application developed under the EUROMEDIA project and currently used by many broadcasters in Europe.

CHAPTER 3

COMPONENTS OF THE PROPOSED FULLY-AUTOMATED SEMANTIC IMAGE CLASSIFICATION SYSTEM

As mentioned in the earlier chapters, semantic classification of images using low-level features is a challenging problem. For an ultimate solution, object-based features should be employed. Unfortunately, extracting the semantic objects from an image requires image segmentation operation and segmentation still remains as an unsolved problem. In the system proposed, the aim is to reach some important generic semantic concepts which can be inferred from the entire image. Therefore, the unsolved problem of segmentation is avoided.

Visual image features that are defined by MPEG-7 standard are used. These features are a product of the long research on the subject and selected by MPEG because of their success and reliability as a result of some experiments. Using MPEG-7 based features for our system not only provides robust tools that are approved by the experts on the subject, but also give the chance to gain experience about a newly emerged standard that will probably constitute the core of future multimedia applications. It should be noted that, MPEG-7 standard is a reliable outcome of the long research that is summarized in Chapter 2.

In order to bridge the gap between the low-level image features and the semantic concepts, *experts* which are classifiers with different structures trained by low-level color and texture features, are utilized. For combining different classifiers and features, two advanced decision mechanisms are proposed [22]. Simulations are conducted on eight different visual semantic classes and performance improvements on accuracy, precision and recall, when compared with single feature and single classifier systems, are observed. The proposed system is implemented as a module in the MPEG-7 compliant multimedia management system, BilVMS, which is developed in TÜBİTAK BİLTEN [23,24,25].

3.1 Low-level Image Features Extraction

The first step towards successful image classification is a good selection among low-level representations (i.e. features). In this research, color and texture descriptors of MPEG-7 [26] are utilized. A total of 4 descriptors are used, which are selected according to their compatibility with the semantic concepts. Two of these descriptors are color-based (color layout and color structure), while the other two (edge histogram and homogeneous texture) are texture descriptors.

3.1.1 Color Layout

The MPEG-7 Color Layout Descriptor (CLD) is a compact and resolution-invariant representation of spatial distribution of colors in an image [26,27]. It is especially recommended for applications that need to be fast and are based on spatial-structure of color.

CLD is obtained by applying DCT transformation on the 2-D array of local representative colors in YCbCr color space where each channel is represented by 8 bits. 8x8 grid, and averaging each of the 3 channels separately for these blocks. A sample divided image is given in Figure 3.1. In the next step, DCT transformation is applied on each of these 8x8 average images. Finally, nonlinear quantization is applied on the coefficients obtained from the previous step. The extraction process is illustrated in Figure 3.2.

Figure 3.1: A sample 8x8 gridded image

Scalable representation of CLD is allowed in the standard meaning that one can select the number of coefficients to use from each channel's DCT output. For each channel, 3, 6, 10, 15, 21, 28 or 64 coefficients can be used. The coefficients are taken from 8x8 arrays in zigzag scan order. Zigzag scan order is illustrated in Table 3.1.

In the proposed system, CLD with 6 Y, 3 Cb and 3 Cr coefficients is extracted and used.

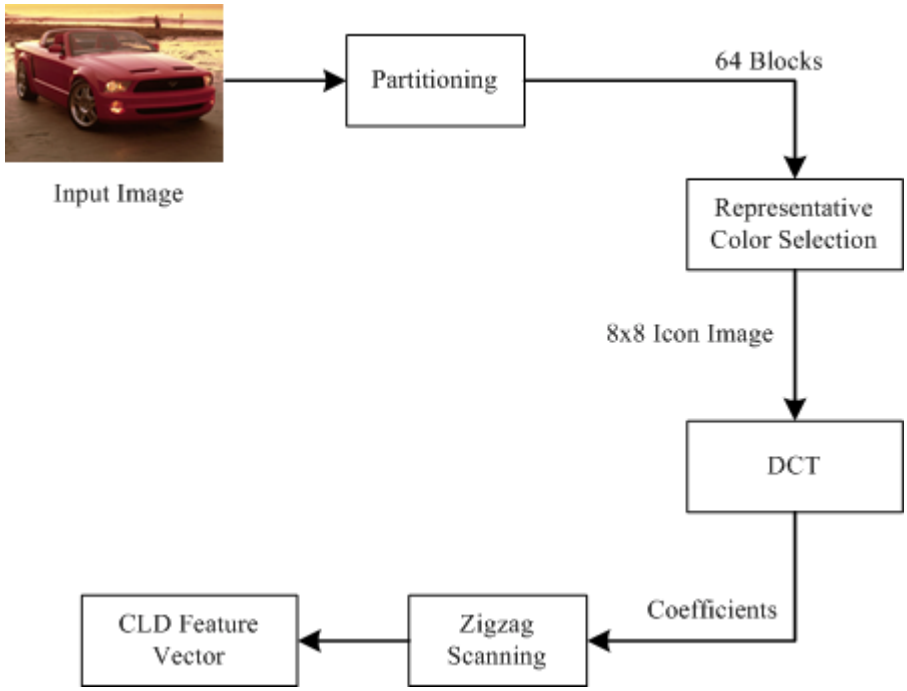


Figure 3.2: Color layout descriptor extraction process

Table 3.1: Zigzag scan order of theDCT coefficients

	i							
j	0	1	5	6	14	15	27	28
	2	4	7	13	16	26	29	42
	3	8	12	17	25	30	41	43
	9	11	18	24	31	40	44	53
	10	19	23	32	39	45	52	54
	20	22	33	38	46	51	55	60
	21	34	37	47	50	56	59	61
	35	36	48	49	57	58	62	63

3.1.2 Color Structure

MPEG-7 Color Structure Descriptor (CSD) specifies both color content (like color histogram) and the structure of this content by the help of a structure element [26,27]. In contrast to a simple image histogram, this descriptor can distinguish between two images in which a given color is present in identical amounts, whereas the structure of the groups of pixels is different. In Figure 3.3, an example illustrating this case is given.

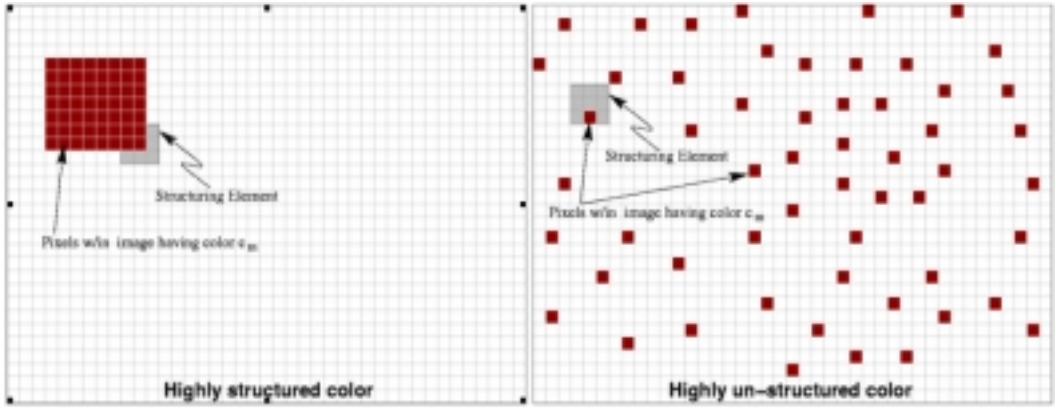


Figure 3.3: Two images having identical color histograms but different color structure descriptors

CSD has a form identical to a color histogram though being semantically different. In an ordinary color histogram, pixels are counted on the whole image in a single session. Therefore, images in Figure 3.3 have identical color histograms. On the other hand, CSD counts the pixels that are inside a shifting structuring element. As a result, images having identical color histograms have different CSDs.

CSD is extracted from an image represented in HMMD color space [26]. Therefore, images that have color spaces other than HMMD shall be converted to this space. Various quantizations can be selected in HMMD color space, as defined in the standard. 256, 128, 64, 32 bin HMMD histograms are allowed. An 8x8 structuring element is used to accumulate the histogram. The accumulation process is illustrated in Figure 3.4. For each unique color that falls inside the structuring element, bins corresponding to these colors are incremented once. For the case of three different colors, as seen in the figure, three of the histogram bins are incremented once. After the accumulation process, follows the non-uniform quantization of the values of the histogram bins according to the statistics of color occurrence, as defined by the standard. This flow is illustrated in Figure 3.5.

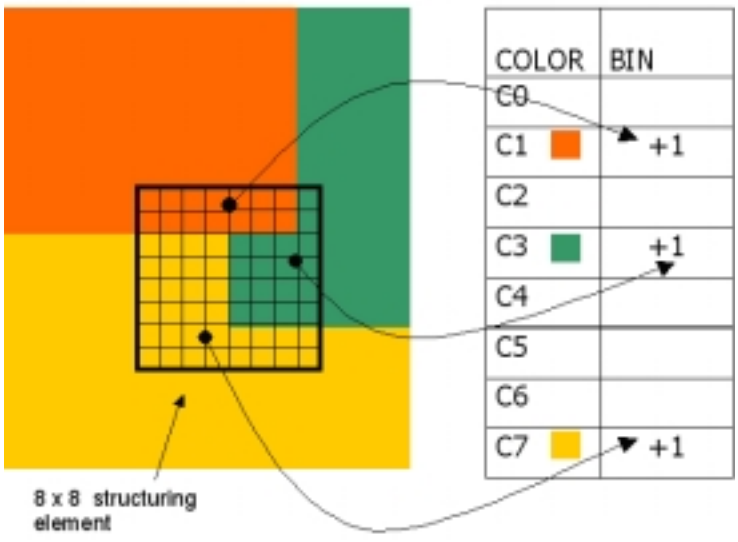


Figure 3.4: Accumulation of histogram bins in color structure descriptor

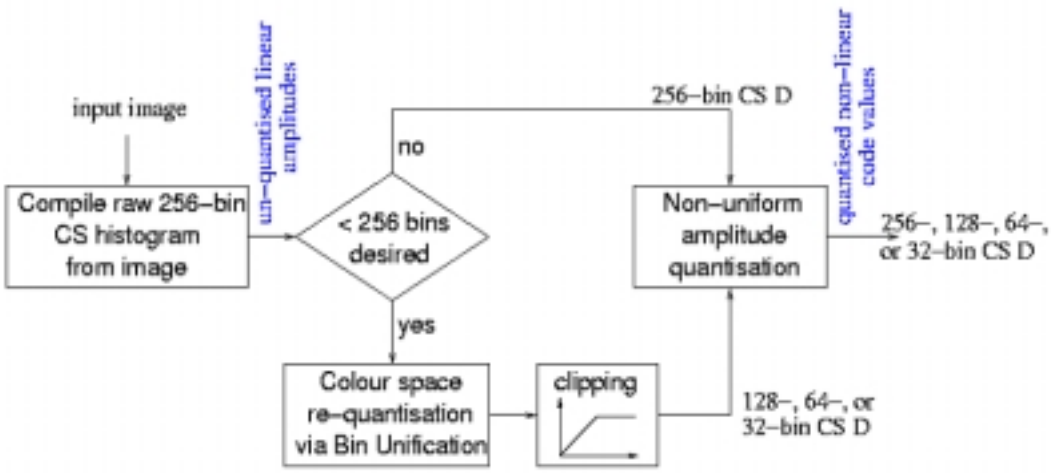


Figure 3.5: Color structure descriptor extraction process

3.1.3 Edge Histogram

Spatial distribution of edges are utilized for image classification in our system by using MPEG-7 Edge Histogram Descriptor (EHD). The EHD represents local edge distribution in an image by dividing the image into 4x4 sub-images and generating a histogram from the edges present in each of these sub-images. Edges in the image are categorized into five types, namely vertical, horizontal, 45° diagonal, 135° diagonal and non-directional edges. In the end, a histogram with 16x5=80 bins is obtained, corresponding to a feature vector having a dimension of 80 [26,27].

As mentioned above, EHD extraction starts with dividing an image into 16 sub-images in a 4x4 grid. These sub-images are indexed according to their locations, as illustrated in Figure 3.6. Next step is performing edge detection inside these sub-images. The filters used in this process are given in Figure 3.7. In the application of filters, if maximum result obtained by the filters exceeds a threshold, an edge with the type of the filter is reported to be found and the

corresponding histogram bin is incremented. The histogram, constructed by the result of this process, is then normalized according to the size of the image.

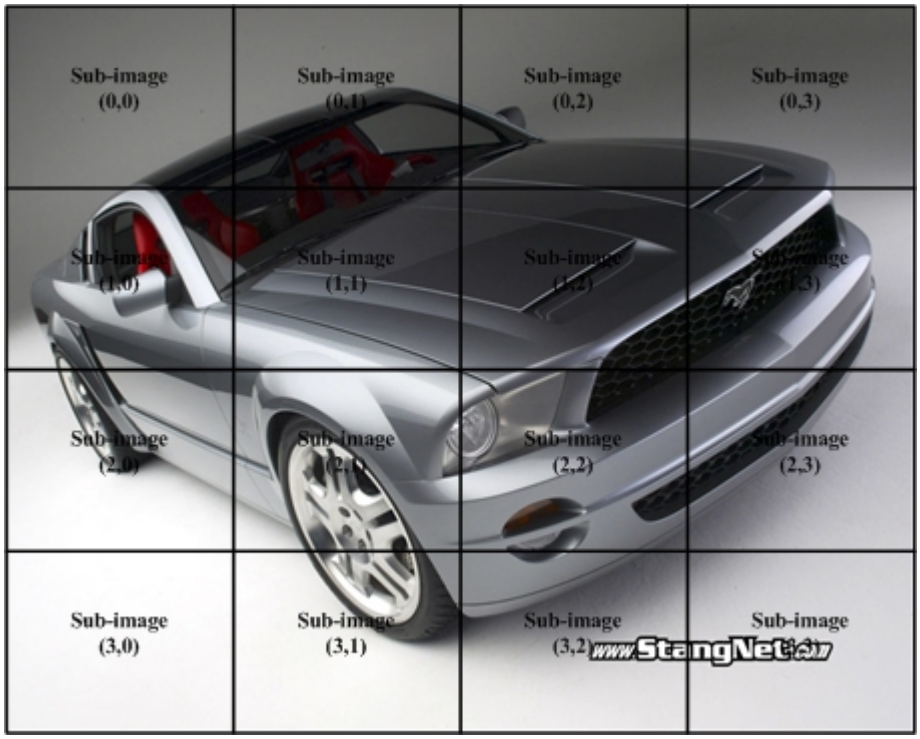


Figure 3.6: A sample image with indexed sub-images

Although not included in the EHD, semi-global and global edge histograms also convey important information. Therefore, during the experimentation process of MPEG-7, these are also computed from the EHD. The recommended way of reaching 13 bin semi-global histogram using the local EHD values is illustrated in Figure 3.8.

Obtaining the global histogram is achieved by a straightforward unification of all the bins related with the same type of edge, resulting in a 5 bin histogram. The dimensions of all three kinds of histograms are given in Figure 3.9.

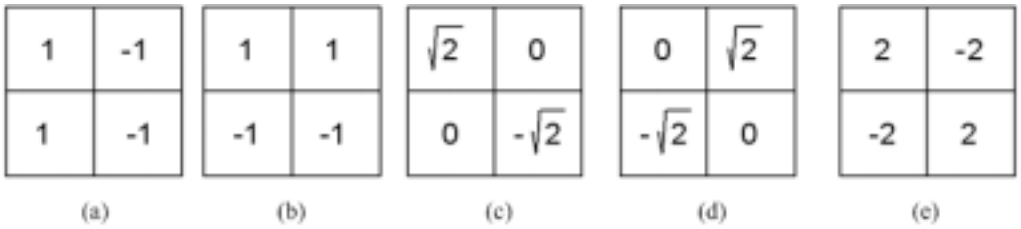


Figure 3.7: Filters used for edge detection (a) vertical, (b) horizontal, (c) diagonal 45º, (d) diagonal 135º, (e) non-directional

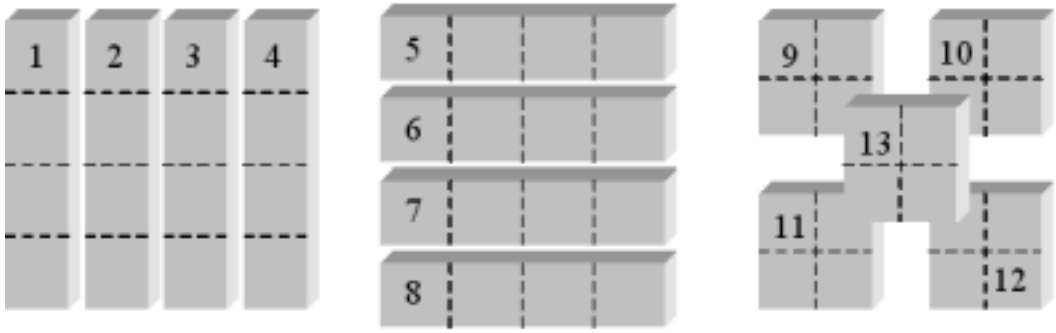


Figure 3.8: Semi-global edge histogram (13 bins for 13 regions)

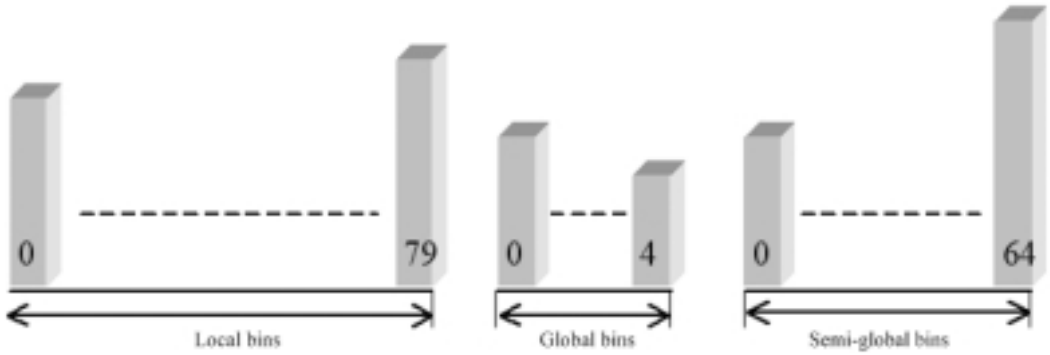


Figure 3.9: Local, semi-global and global histogram bins

3.1.4 Homogeneous Texture

MPEG-7 Homogeneous Texture Descriptor (HTD) characterizes the region texture by mean energy and energy deviation from a set of frequency channels. The channels are modeled by Gabor functions and the 2-D frequency plane is partioned into 30 channels. In order to construct the descriptor, the mean and the standard deviation of the image in pixel domain is calculated and combined into a feature vector with the mean and energy deviation computed in each of the 30 frequency channels. As a result, a feature vector of 62 dimensions is extracted from each image [26,27].

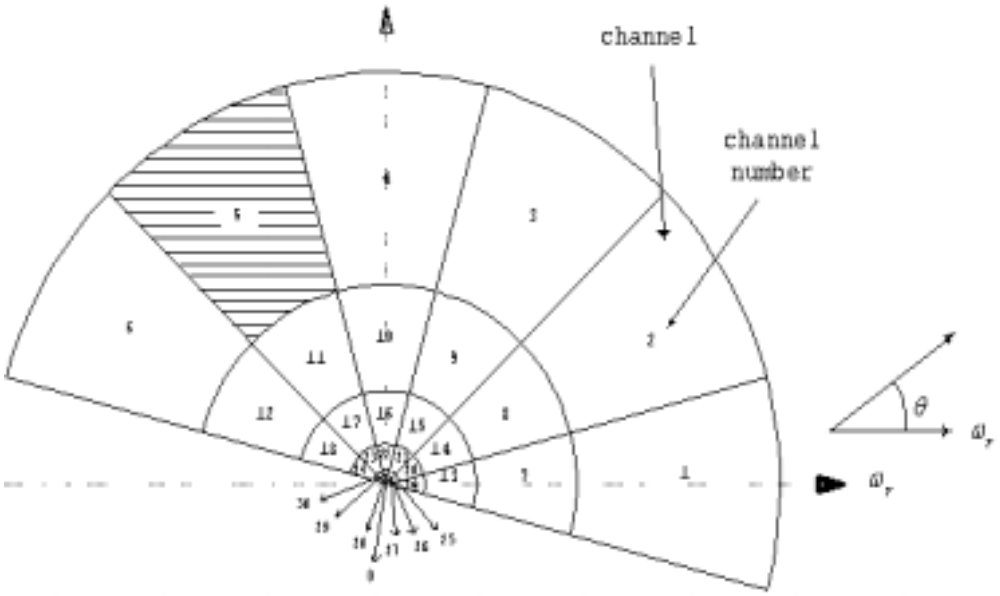


Figure 3.10: Frequency channels on the 2-D frequency plane

Channels spanning the 2-D frequency plane, as illustrated in Figure 3.10, are indexed by 5 radial and 6 angular indexes and are defined by the following formula:

$$G_{P_{S,r}}(\omega,\theta) = \exp\left[\frac{-(\omega-\omega_s)^2}{2\sigma_{\rho_s}^2}\right]\cdot\exp\left[\frac{-(\theta-\theta_r)^2}{2\sigma_{\theta_r}^2}\right]$$

In the normalized frequency space ($0 \le \omega \le 1$), the channels illustrated in Figure 3.10 are equally spaced in the angular direction, such that $\theta_r = 30° \times r$, while spacing in the radial direction is in octave scale, such that the bandwidth is defined by $B_s = B_0 \cdot 2^{-s}$ where B_0 is 0.5. In the above formula, s is the radial index and r is the angular index, where s$\in$ {0,1,2,3,4}and r$\in$ {0,1,2,3,4,5}.

σ_{θ_r} and σ_{ρ_s} are the standard deviations of the Gaussian distribution and are defined by $\sigma_{\theta_r} = 15°/\sqrt{2\ln 2}$ and $\sigma_{\rho_s} = B_s/2\sqrt{2\ln 2}$. Note that the standard deviation along the radial direction is constant.

The mean energy of each channel is computed by the formula $e = \log_{10}[1+p]$ where,

$$p = \int_{\omega=0^+}^{1}\int_{\theta=(0°)^+}^{360°}[G_{P_{S,r}}(\omega,\theta)\cdot P(\omega,\theta)]^2$$

On the other hand, standard deviation is calculated from each channel by $d = \log_{10}[1+q]$, where q is defined as:

$$q = \sqrt{\int_{\omega=0^+}^{1}\int_{\theta=(0°)^+}^{(360°)^+}\left\{\left[G_{P_{S,r}}(\omega,\theta)\cdot P(\omega,\theta)\right]^2 - p\right\}^2}$$

The results of these mean and standard deviation calculations are to be coded from right to left and from outside to inside as seen in Figure 3.10.

3.2 Classifiers

In this thesis, four popular classifiers are utilized, which are Support Vector Machine, Nearest Mean, Bayesian Gaussian Plug-In and K-Nearest Neighbors. Binary classification is performed by some experts, which are obtained via training these classifiers with in-class and informative out-of-class samples. These classifiers are selected due to their distinct natures of modeling a distribution. For distance-based classifiers (i.e. Nearest Mean and K-Nearest Neighbor), special distance metrics compliant with the nature of the MPEG-7 descriptors are also utilized [26]. Since the outputs of the classifiers are to be used in combination, modifications are achieved on some of them to convert uncalibrated distance values to calibrated probability values in the range [0,1]. All of these modifications are explained in detail along with the structure of the classifiers in the following subsections.

3.2.1 Support Vector Machines (SVM)

SVM is a newly introduced machine learning technology that is based on strong theoretical foundations [28,29,30]. It performs classification between two classes by finding a decision surface that is based on the most informative points of the training set. Its empirical success has been proved by experiments in many different areas including handwritten digit recognition, text classification, face recognition and object recognition [31,32,33].

The main reason behind the success of SVM is the way that it handles the "risk" concept. Although other classical classifiers try to classify the training set with minimal errors and therefore reduce the *empirical risk*, SVM can sacrifice

from training set performance for being successful on yet-to-be-seen samples and therefore reduces *structural risk* [30]. Briefly, SVM constructs a decision surface between the samples of two classes, maximizing the margin between them differing from the other classifiers, as illustrated in Figure 3.11.

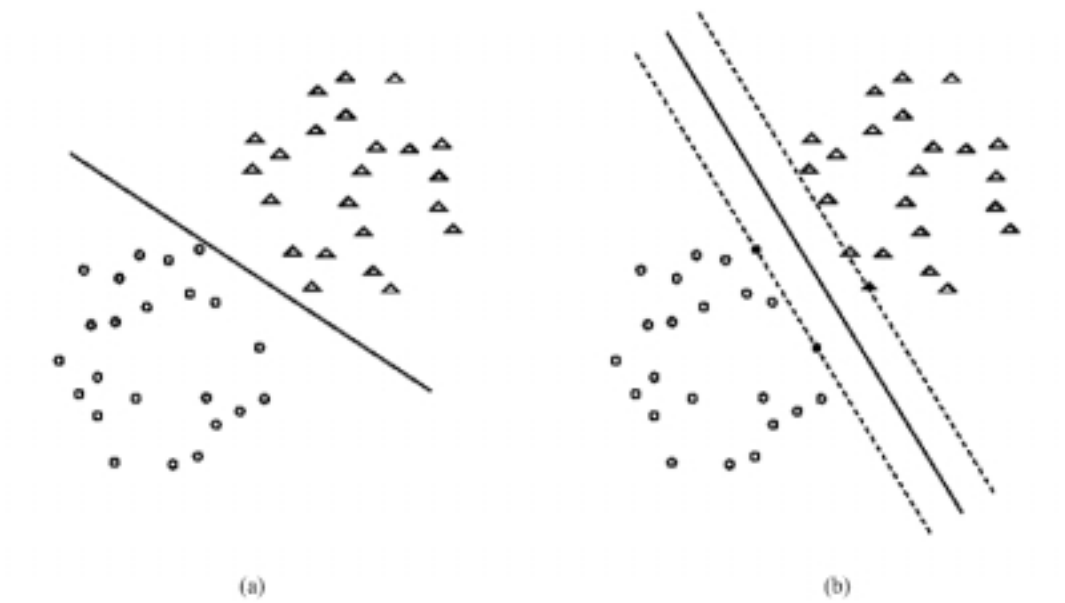


Figure 3.11: (a) boundary obtained by an ordinary classifier, (b) boundary obtained by SVM

Assume a training data set $\{x_1,\ldots,x_n\}$ is given consisting of vectors in a space $X \subseteq R^d$, and their labels $\{y_1,\ldots,y_n\}$ where $y_i \in \{-1,+1\}$. SVM projects the data in the original feature space X to a higher dimensional space *F* by using a kernel operator *K*. Then, in this new induced feature space *F*, the hyperplane providing the maximum margin, which is also called optimal separating hyperplane (OSH) is sought [30]. If the training data is separable, SVM separates the classes and maximizes the margin between them. If the training data is not separable, the solution is a trade-off between the largest margin and the lowest

number of training errors. This trade-off is controlled by a regularization parameter. Samples lying on one side of the hyperplane is labeled as 1, while samples on the other side are labeled as -1. Support vectors that gave the name of the classifier and specify the references for the boundary, are selected from the training instances that determine the boundary.

The functional representation of the classifier explained above is $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$. Input x is classified as 1, if $f(x) \geq 0$, and as -1 otherwise. The kernel function K is usually time separable and can be written as $K(u,v) = \Phi(u) \cdot \Phi(v)$ where $\Phi : \mathrm{X} \rightarrow F$ and "." signifies the inner product operation. Therefore, the classifier function can be expressed as [31]:

$$f(x) = w \cdot \Phi(x), \text{ where } w = \sum_{i=1}^{n} \alpha_i \Phi(x_i).$$

Although K is not seen in this representation, training data is implicitly projected by the kernel into a higher dimensional space. By choosing the type of the kernel used here, one can project the data into different induced feature spaces F in which separating hyperplanes correspond to more complex boundaries in the original feature space X.

There are many different kernel types, such as polynomial, Gaussian, radial basis function and neural network that are used in SVMs [29]. In this research, a second degree polynomial kernel, which can be represented by $K(u,v) = (u \cdot v + 1)^2$ is used. Note that training data feature vectors that are transformed by the kernel should be normalized such that the modulus of them

($\|x_i\|$) are constant or at least the modulus of the vectors in the induced space ($\|\Phi(x_i)\|$) are constant.

In the next step, SVM computes $\alpha_i's$ that correspond to the OSH. In fact, OSH is a member of the version space *V,* that consists of all the hyperplanes successful in separating the data. In other words, every member of the version space satisfies $f(x_i) > 0$, if $y_i = 1$ and $f(x_i) < 0$, if $y_i = -1$. This fact can be explained by a more formal definition as follows:

The set of all possible hyperplanes is defined by:

$$\mathrm{H} = \left\{ f(x) = \frac{w \cdot \Phi(x)}{\|w\|}, w \in W \right\}$$ where W is the parameter space and simply equal to F.

The version space consisting all of the separating hyperplanes can then be represented as:

$$V = \{ f \in \mathrm{H} \quad \forall i \in \{1,...,n\} \quad y_i f(x_i) > 0 \}$$

If one expresses *f(x)* in terms of *w*, then one can redefine *V* as:

$$V = \{ w \in W \quad \|w\| = 1, \quad y_i (w \cdot \Phi(x_i)) > 0, \quad \forall i \in \{1,...,n\} \}$$

Note that the version space *V* defined above can only exist, if the training data is linearly separable in the induced feature space. Otherwise, it is meaningless to search for an OSH. However, this restriction can be overcome by modifying the kernel used so that the data in the induced feature space become separable. This can be achieved by redefining the kernel with the addition of a regularization constant, i.e. $K(x_i, x) \rightarrow K(x_i, x) + v$.

In order to illustrate the selection of the hyperplane, the best way is to utilize the duality between the induced feature space *F* where the training data belongs and the parameter space *W,* where the hyperplanes belong. Points in *W* correspond to hyperplanes in *F* and points in *F* correspond to hyperplanes in *W*. This duality leads us to the result that if the points in *W* restricts the points in *F*, the same must apply in the converse case. This means that each point in the training set defines a hyperplane that restricts the region to which *w* may belong. This duality can be observed from the decision formula with a small rearrangement:

$$y_i(w \cdot \Phi(x_i)) = w \cdot y_i\Phi(x_i) > 0$$

As can be seen in the above formula, $y_i\Phi(x_i)$ can be considered as a normal vector in *W,* which defines a half-space in the above inequality.

SVM finds the hyperplane that maximizes the distance of the closest point to the boundary, subject to the constraint $\|w\| = 1$. Notice that this constraint ensures that version space is on the surface of a unit hypersphere in *W*. This hypersphere is intersected by the hyperplanes that are defined in *W* by the training data in *F*. Using SVM, the center of this hypersphere, which has the largest radius, and whose surface does not intersect with the training data defined hyperplanes, is tried to be found. This center, as already been emphasized, lies on the unit hypersphere in *W*. The hyperplanes tangent to this hypersphere correspond to support vectors and the radius of the hypersphere is the margin of the SVM. These observations are illustrated by visualizations in Figure 3.12.

The search for the maximal margin hyperplane is actually an optimization problem. Although the problem is stated clearly, solving it for large learning tasks

involving many training examples is a real challenge. In this research, the SVM learner that is proposed in [34] is utilized. This implementation is called SVM^{light} [35] and is developed using the algorithmic and computational results in [34] and makes large-scale SVM learning practical.

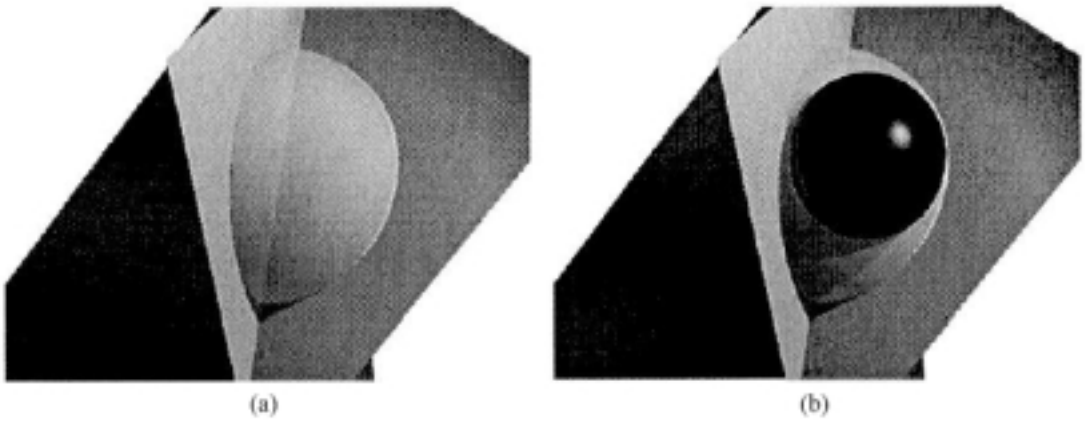


Figure 3.12: (a) Unit hypersphere W that version space resides in, (b) Largest radius hypersphere whose center is on the unit hypersphere

SVM classifier takes as input a data vector and finds out at which side of the classifying hyperplane it resides, as well as its distance to the decision surface. If the input sample resides at the in-class half of the space divided by the hyperplane, then the distance value is preceeded by a plus "+" sign, otherwise the distance value is preceeded by a minus "-" sign. This output format, though being meaningful for the specific context, is not appropriate for combination with other classifier outputs. In order to obtain a calibrated posterior probability value, a simple logistic link function method, proposed by Wahba [36], is utilized. Using this method, posterior probability of a sample to be in-class is computed from its distance to the boundary by the following formula:

$$\mathrm{P}(in-class \mid x) = \frac{1}{1+e^{-f(x)}}$$

In this formula, *f(x)* is the uncalibrated output of SVM, which is the signed distance of the input vector from the decision surface. It should also be noted that this conversion conforms with the important requirements for post processing of classifier outputs [37], which should be taken into account in order not to experience degradations in the classifier performance.

3.2.2 Nearest-Mean Classifier

Nearest mean classifier calculates the centers of in-class and out-of-class training samples and then assigns the upcoming samples to the closest center. This classifier is included into this research due to its ability to model compact distributions effectively.

This classifier, like SVM, gives distance values at its output. These two distance values, which are measured between the mean of each distribution and the input vector, should be modified to produce a calibrated posterior probability value. A common method used for distance-based K-NN classifiers is adapted to this case [38]. According to this method, distance values are mapped to posterior probabilities by the formula,

$$\mathrm{P}(in-class \mid x) = \frac{\frac{1}{d_{m_{in-class}}}}{\frac{1}{d_{m_{in-class}}} + \frac{1}{d_{m_{out-of-class}}}}$$

where $d_{m_{in-class}}$ and $d_{m_{out-of-class}}$ are distances of the input x from the in-class and out-of-class training set means respectively.

Another measure to increase the confidence of the classification, which is also proposed in [38], is applied when the probability obtained from the previous

step is below an ambiguity threshold (0.55 is selected as a result of the pre-experiments). This second measure computes the probability values by,

$$\mathrm{P}(in-class \mid x) = \frac{\mathrm{N}_i}{\mathrm{N}}$$

where N_i is the number of in-class training samples whose distance to the mean is greater than x, and N is the total number of in-class training samples. By these steps, uncalibrated classifier outputs are not only converted to calibrated probabilities in [0,1], but also a more effective nearest mean classifier is obtained using the underlying details of the training data.

3.2.3 Bayesian Gaussian Plug-in Classifier

Bayesian Gaussian Plug-in classifier fits multivariate normal densities to the distribution of the training data. In binary classification, two class conditional densities representing in-class and out-of-class training data are obtained, as a result of this process. Bayesian decision rule is then utilized to find the probability of the input to be a member of the semantic class [39].

Multivariate normal densities that are to be fitted to the training data are in the form,

$$\mathrm{P}(x) = \frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

where x is a d dimensional input feature vector, μ is a d dimensional mean vector and $\sum$ is a dxd covariance matrix. If we have n training samples then μ is computed by,

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{3.1}$$

Covariance matrix is then computed using the above result by the formula:

$$\Sigma = \frac{1}{n}\sum_{k=1}^{n}(x-\mu)(x-\mu)^T \tag{3.2}$$

where x and μ are dx1 vectors. After class conditional probability distribution functions of in-class and out-of-class training data are obtained by Eqns. (3.1) and (3.2), models representing each class contained in two functions, *P(x | in-class)* and *P(x | out-of-class)*, can be obtained.

The next step towards the final decision about an input data is utilizing Bayesian decision rule [39], whose class conditional densities are obtained in the previous step. According to the Bayesian decision rule, posterior probability of being in-class is computed with the formula,

$$\mathrm{P}(in-class|\, x) = \frac{\mathrm{P}(x\,|\,in-class)\mathrm{P}(in-class)}{\mathrm{P}(x\,|\,in-class)\mathrm{P}(in-class)+\mathrm{P}(x\,|\,out-of-class)\mathrm{P}(out-of-class)}$$

for an input sample. Throughout the research, the *apriori* probability values for in-class and out-class are fixed as 0.5 and therefore neglected as:

$$\mathrm{P}(in-class\,|\,x) = \frac{\mathrm{P}(x\,|\,in-class)}{\mathrm{P}(x\,|\,in-class)+\mathrm{P}(x\,|\,out-of-class)}$$

This classifier is again successful at modeling natural distributions that have a characteristic compaction around a center. The main disadvantage of this classifier is the degradation in modeling performance with the increasing feature dimension.

3.2.4 K-Nearest Neighbor Classifiers (K-NN)

K-NN classifiers are known to be successful, especially while capturing important boundary details that are too complex for all of the previously mentioned

classifiers. Due to this property, they can model sparse distributions with a relatively high accuracy.

Generally, the output of these classifiers are converted to probability, except for K=1 case, by the following formula:

$$\mathrm{P}(w_{i|} \mid x) = \mathrm{K}_i / \mathrm{K}$$

where K_i represents the number of nearest neighbors from class w_i and K is the total number of nearest neighbors, taken into consideration. This computation, although quite simple, underestimates an important point about the location of the test sample relative to in-class and out-of-class training samples. Therefore, instead of the above method, a more complex estimation method is adapted and utilized in this research:

$$\mathrm{P}(in-class \mid x) = \frac{\sum_{y_i} \frac{1}{d(x, y_i)}}{\sum_{j=1}^{k} \frac{1}{d(x, y_j)}}$$

where y_i shows in-class nearest neighbors of the input and y_j represents all k-nearest neighbors of the input.

Although, this estimation provides a more reliable probability output, it is observed that applying another measure to the test samples with probabilities obtained by the above formula that are below a threshold, improves the result further. This measure utilizes the relative positions of training data among each other [38]. This metric is the sum of the distances of each in-class training sample to its k in-class nearest neighbors and obtained by,

$$g(x) = \sum_{i=1}^{k} d(x, y_i) \qquad y_i : i^{th} \text{ in-class nearest neighbor}$$

After the computation of this value for each training sample, and the input test sample, the final probability is obtained by

$$\mathrm{P}(in-class \mid x) = 1 - \frac{N_i}{N}$$

where N_i is the number of in-class training samples with *g(x)* value smaller than the input test sample and N is the number of all n-class training samples. In this way, significant improvements are achieved in 3-NN, 5-NN, 7-NN and 9-NN classifier performances.

For special case of 1-NN (Nearest Neighbor), the explained conversion technique is not applicable. Therefore, for 1-NN classifier the same probability estimation technique as that employed in the nearest mean classifier is applied.

3.3 Expert Combination Strategies

In the area of pattern recognition, many different schemes have been developed in order to achieve the best possible classification performance by combining the experts [40,41,42]. *Experts* are instances of classifiers with distinct natures working on distinct feature spaces, and their combination has been a popular research topic for years. Latest studies have provided mature and satisfying schemes for expert combination [43]. Although many of these schemes developed specifically for a task at hand, and making use of heuristics of their applicable domain, there also exist some fundamental rules from which most of the other schemes are devised.

Six fundamental expert combination rules are used in this research and they are considered to be the most reliable ways of using the complementary information offered by different classifier structures. These combination rules,

namely product, sum, max, min, median and majority vote, and the relations among them have been theoretically analyzed in depth, and many experiments have proved that, in many situations, they outperform the single experts that are included in the combination [43].

According to the utilization of the intermediate classification results, combination schemes are divided into three main categories that are abstract level, rank level and measurement level combinations [37,41]. The last of these three is known to be the one conveying the highest information from the intermediate step. Measurement level combination, which is the case in the first five of the combination rules below, uses the outputs of expert functions directly. Abstract level combination is the category where only the decision labels are used, and to which the last rule, namely majority vote, belongs.

Notwithstanding the successful results achieved in combination experiments using these rules, the reasons leading to the superiority of a scheme over the others are not obvious. Therefore, special circumstances that this success will repeat, have not been adequately understood yet, even for the most fundamental ones [44].

In all the rules, a priori probabilities are assumed as 0.5 and the decision is made by the following formula:

$$\mathrm{P}(in-class \mid X)=\frac{\mathrm{P}_1}{\mathrm{P}_1+\mathrm{P}_2} \qquad \mathrm{P}(out-of-class \mid X)=\frac{\mathrm{P}_2}{\mathrm{P}_1+\mathrm{P}_2} \qquad (3.3)$$

In this formula, P_1 is the combined output of experts about the posterior probability of the sample X belonging to the semantic class, while P_2 is the posterior probability of the sample not belonging to the semantic class. This

formula generates probability outputs and normalizes the scale distortion of them into the range [0,1]. The final decision is given according to the Bayes' Rule, if posterior probability of the sample being a member of the class is equal to or above 0.5, then the sample is assigned as in-class, else it is assigned out-class. P_1 and P_2 are obtained by using the combination rules, namely *product rule*, *sum rule*, *max rule*, *min rule*, *median rule*, and *majority vote*. The last four of these rules are derived from the first two rules which have particular statistical assumptions behind their developments. These assumptions are made in order to reach practicable rules from the Bayesian decision rule, which depends on joint probability density functions that are difficult to infer.

3.3.1 Product Rule

Product rule is developed using the Bayesian decision theory, under the assumption that densities used by each expert are conditionally statistically independent. The cause of this independence may be the use of different representations of data or different distribution models. In this rule, R experts are combined as follows:

$$\mathrm{P}_1 = \prod_{i=1}^{R} \mathrm{P}_i(in-class \mid X) \qquad \mathrm{P}_2 = \prod_{i=1}^{R} \mathrm{P}_i(out-of-class \mid X)$$

Note that, in the above formulas terms depending on a priori probabilities are omitted, due to the equal a priori probabilities assumption.

This rule is known to be a severe rule of combining experts, since one of the experts can severely affect the combination result by outputting a probability close to zero.

3.3.2 Sum Rule

This rule is developed under assumptions which are stricter than the product rule. In addition to the conditional independence assumption in the product rule, sum rule assumes that probability distribution will not deviate significantly from the a priori probabilities. Although this is a strong assumption that may not be true if the noise level of the observations is not high enough, it still provides an adequate and useful approximation in many situations.

According to the sum rule, the posterior probabilities used in Eqn. 3.3 are calculated as,

$$\mathrm{P}_1 = (1-R)\mathrm{P}(in-class) + \sum_{i=1}^{R} \mathrm{P}_i(in-class \mid X) \tag{3.4}$$

$$\mathrm{P}_2 = (1-R)\mathrm{P}(out-of-class) + \sum_{i=1}^{R} \mathrm{P}_i(out-of-class \mid X) \tag{3.5}$$

In this rule, equal a priori probabilities do not cancel and therefore they are not omitted. However, the output of the sum rule may not be in the appropriate probability interval and therefore values less than zero and greater than one are corrected as zero and one, respectively. Sum rule has almost inverse properties with the product rule. It is robust against outliers and may be viewed as an average over the results, as mentioned above.

3.3.3 Max Rule

This rule is derived from the sum rule, under the assumption of equal a priori probabilities. Outputs are computed in this rule by,

$$\mathrm{P}_1 = \max_{i=1}^{R} \mathrm{P}_i(in-class \mid X) \qquad \mathrm{P}_2 = \max_{i=1}^{R} \mathrm{P}_i(out-of-class \mid X)$$

3.3.4 Min Rule

Min rule is derived from the product rule and assumes equal a prior probabilities like the max rule. Calculation of the probabilities are by means of the following:

$$\mathrm{P}_1 = \min_{i=1}^{R} \mathrm{P}_i(in-class \mid X) \qquad \mathrm{P}_2 = \min_{i=1}^{R} \mathrm{P}_i(out-of-class \mid X)$$

3.3.5 Median Rule

Sum rule, under equal a priori assumption, is similar to averaging posterior probabilities. However, in order to avoid any disturbance of average value from the probable outliers, a robust estimate of the mean, namely median is used.

Combined estimation is calculated by the median rule in a quite similar way to max and min rules:

$$\mathrm{P}_1 = \underset{i=1}{\overset{R}{median}}\, \mathrm{P}_i(in-class \mid X) \qquad \mathrm{P}_2 = \underset{i=1}{\overset{R}{median}}\, \mathrm{P}_i(out-of-class \mid X)$$

Note that in the case of even number of experts, since median does not exist, it is replaced by mean.

3.3.6 Majority Vote

Majority vote is derived from the sum rule under equal a priori probability assumption. In this rule, instead of measurement level, abstract level expert combination is performed. P_1 is assigned the number of experts having in-class probabilities above 0.5 and therefore P_2 is equal to the total number of experts minus P_1. Both of these values are then normalized by the total number of experts that is equal to P_1 plus P_2. This can be represented by the formulas below:

$$\mathrm{P}_1 = \frac{N_1}{N_1 + N_2} \qquad \mathrm{P}_2 = \frac{N_2}{N_1 + N_2}$$

where N_1 is the number of experts voting as in-class, and N_2 is the number of experts voting as out-of-class.

3.4 Advanced Decision Mechanisms

In the proposed semantic image classification system, in order to improve the classification performance and stability, advanced decision mechanisms are implemented. These mechanisms utilize the expert combination strategies explained in the previous section to combine the experts having classifier structures defined in Section 3.2, and are trained by low-level visual image features defined in Section 3.3.

The decision mechanisms proposed here, namely Single Feature Combination (SFC), Multiple Feature Direct Combination (MFDC), and Multiple Feature Cascaded Combination (MFCC), constitute the core of the experiments conducted throughout this research [22].

3.4.1 Single Feature Combination (SFC)

The most simple decision mechanisms is single feature combination, which is merely a realization of the expert combination strategies for a single low-level feature. In this setting, experts having distinct classifier structures are trained by a single low-level feature. After the training process, an input vector to be classified is first introduced to these experts. Then, the calibrated probability outputs of the experts are merged into a single calibrated posterior probability value using one of the combination strategies defined in the previous section. SFC mechanism is illustrated in Figure 3.13.

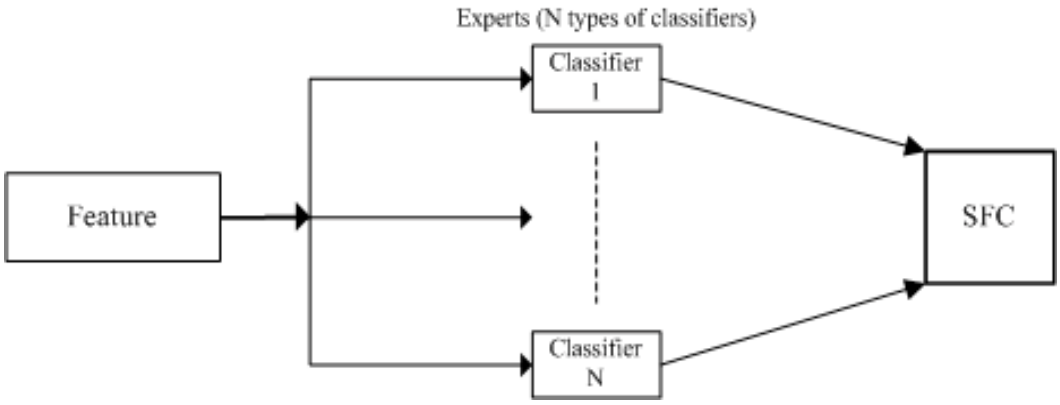


Figure 3.13: Single Feature Combination (SFC)

3.4.2 Multiple Feature Direct Combination (MFDC)

For the cases where more than one low-level visual image feature is required for defining a visual semantic concept, experts trained with different visual features should be combined. MFDC combines output of single experts that are trained by multiple features in a single step by using the expert combination strategies.

MFDC is illustrated by a diagram in Figure 3.14. In the diagram, a case where M features and N different classifier structures are used, is visualized. As seen in the figure, for each of the M features, an ensemble of N experts are generated, resulting in a total of MxN experts.

3.4.3 Multiple Feature Cascaded Combination (MFCC)

Another mechanism that is used to merge the information coming from experts trained by multiple features is the multiple feature cascaded combination (MFCC). In this setting, SFC outputs obtained by combining multiple experts trained by a single feature in the first step are utilized in a second step. MFCC mechanism is illustrated in Figure 3.15.

In Figure 3.15, this mechanism is visualized for M feature and N classifier structure case. In the first step, each feature vector is presented to the ensemble of classifiers trained for each feature separately. The outcome of this process is MxN calibrated probability values that are appropriate for combination. Each ensemble of N experts trained by the same low-level feature is then merged to give M probability values which are nothing but the SFC outputs of each feature. These outputs are then further merged using one of the rules in Section 3.3 to obtain the MFCC result.

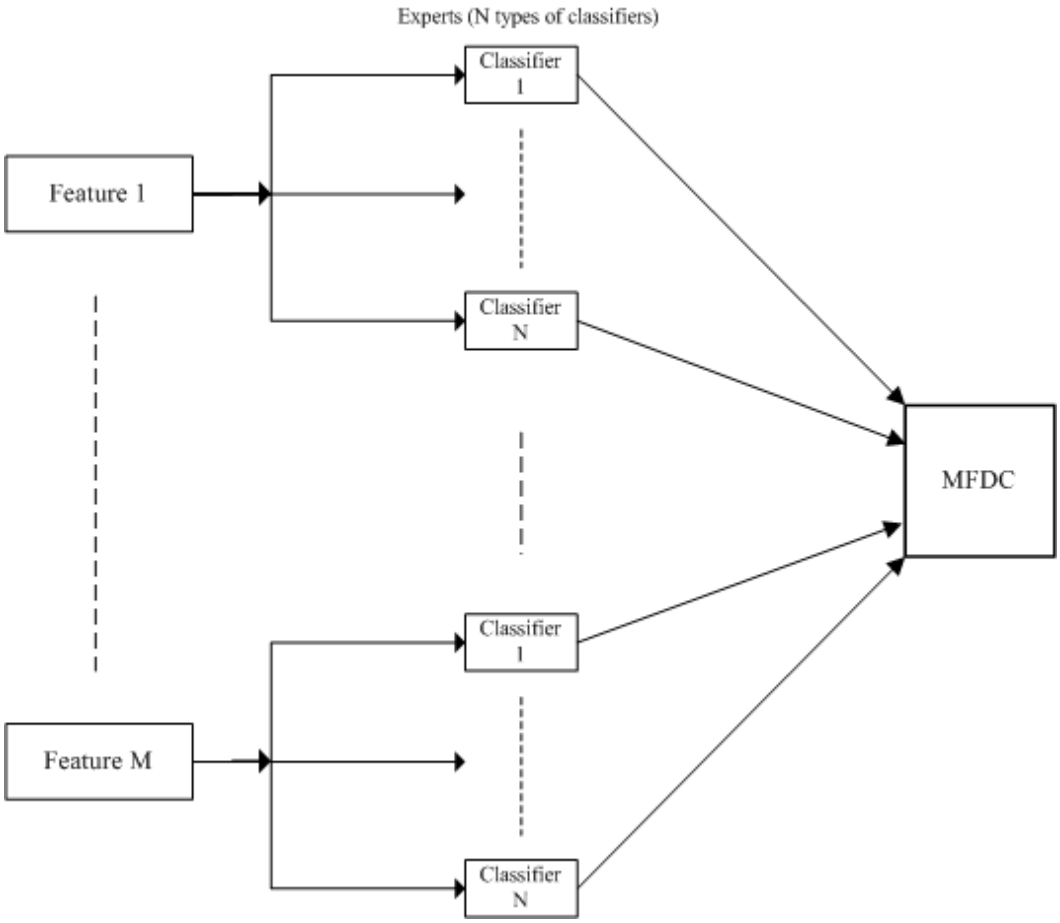


Figure 3.14: Multiple Feature Direct Combination (MFDC)

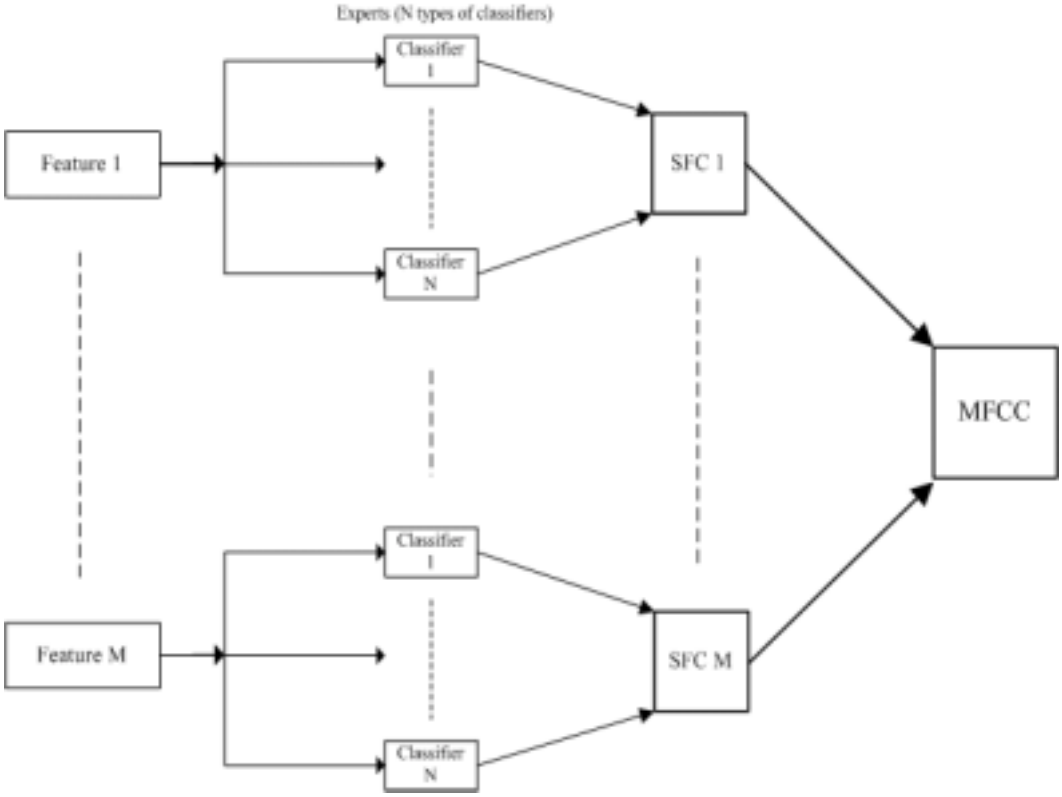


Figure 3.15: Multiple Feature Cascaded Combination (MFCC)

CHAPTER 4

SIMULATIONS

The system, whose components are explained in the preceding sections, is tested and results are presented in this chapter. A total of 1600 images, collected from various resources and having different resolutions are utilized during training and test phases. Eight different visual semantic classes are classified using both single experts and their combinations. The flow of the system is illustrated in Figure 4.1. The illustrated system is implemented as a C++ program, which allows the selection of classifier types, low-level image features and decision mechanisms. This program is also a module of the MPEG-7 compliant video management system, BilVMS, which is developed in TÜBİTAK BİLTEN [23,24].

4.1 Semantic Classes and Representative Features

All of the eight semantic classes that are classified in this research have the common property of being convenient to be inferred from low-level visual features of the entire image. In other words, the characteristics of these classes are usually significant in the entire image.This property, as already mentioned, enables us to avoid the need for segmentation of the images.

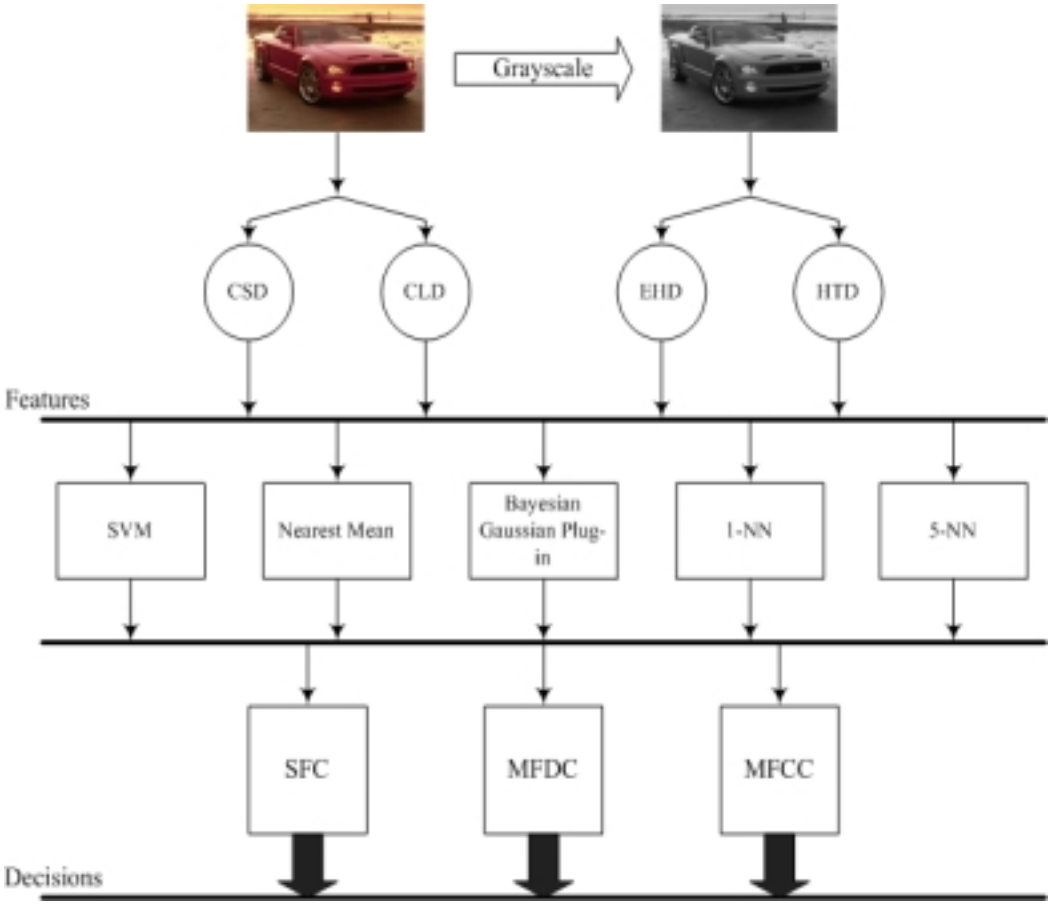


Figure 4.1: System Flow

A list of classes and their utilized low-level visual features are given in Table 4.1. Note that three of these classes are defined by a single feature, while five of them are defined with multiple features. These features are selected according to the results of some preliminary experiments which are not documented here. During those experiments, the abilities of the MPEG-7 low-level visual descriptors to represent the semantic details in images are analyzed. As a result, each semantic entity is related with some low-level features that proved to capture the characteristics best in the preliminary experiments. In Figure 4.2, some typical images from each of the semantic classes are given.

Table 4.1: Semantic classes and corresponding low-level features

Semantic Class	**Low-Level Features**
Football	Color Layout
Indoor	Edge Histogram
Crowd	Homogeneous Texture
Sunset-Sunrise	Color Layout, Color Structure, Edge Histogram
Sky	Color Layout, Color Structure, Homogeneous Texture
Forest	Color Structure, Edge Histogram, Homogeneous Texture
Sea	Color Layout, Homogeneous Texture
Cityscape	Color Structure, Edge Histogram, Homogeneous Texture

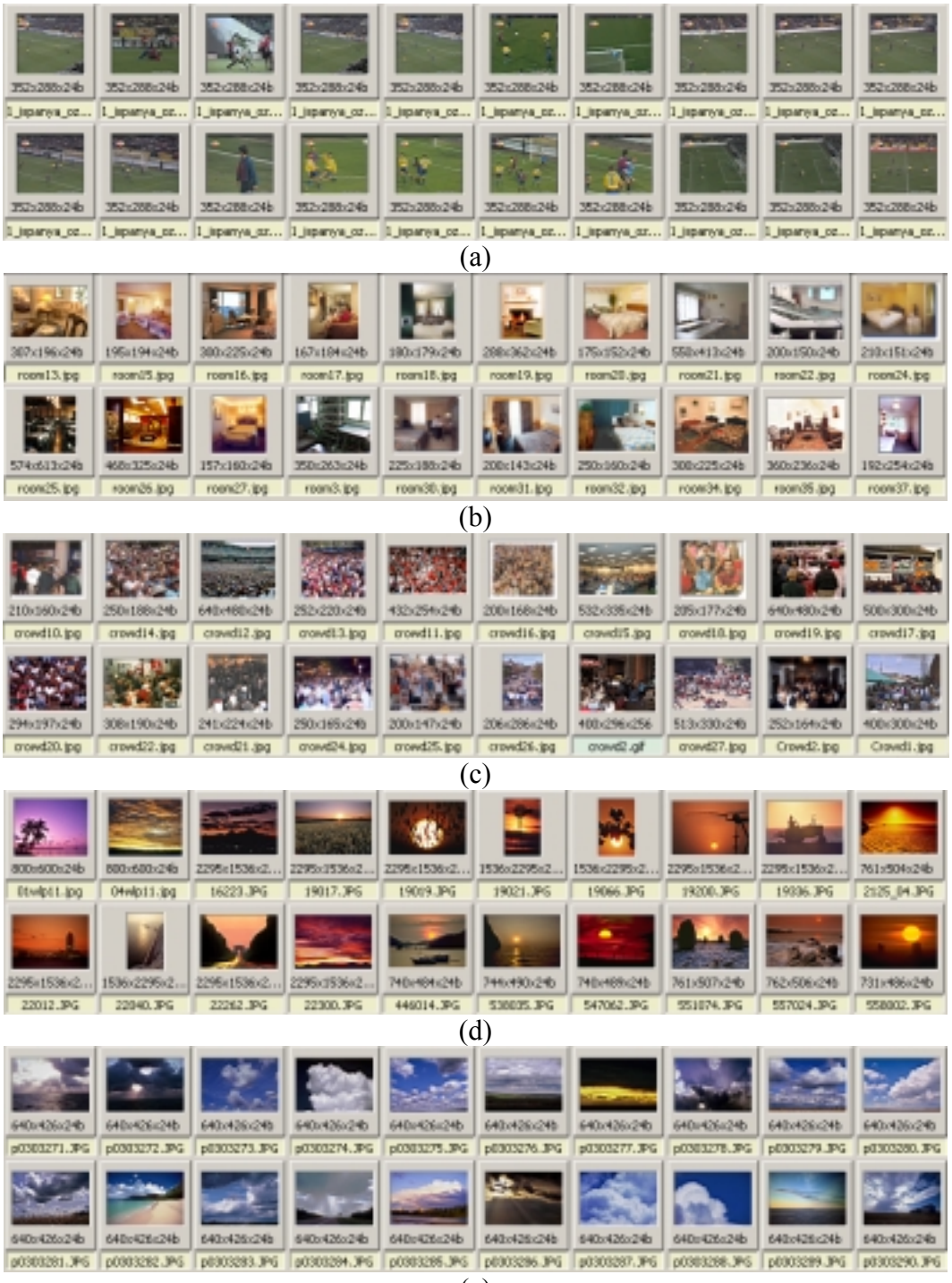
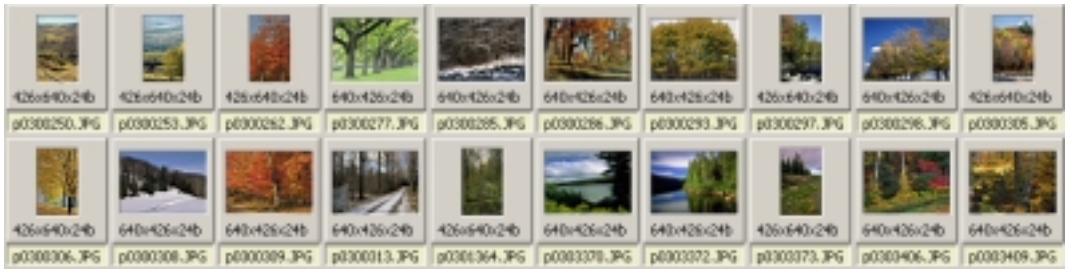
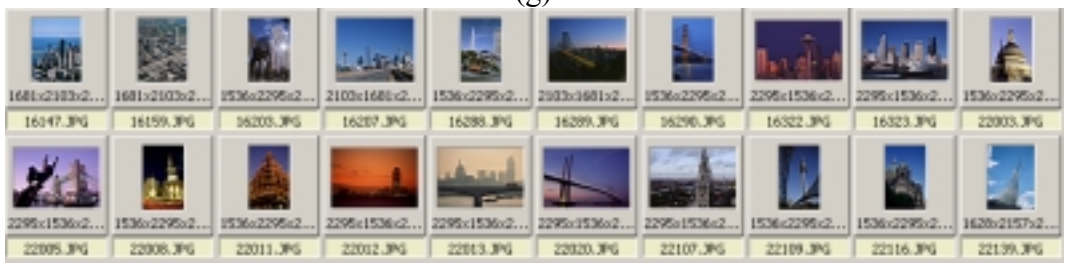


Figure 4.2: Typical images from semantic classes (a) football, (b) indoor, (c) crowd, (d) sunset-sunrise, (e) sky



(f)

(g)



(h)

Figure 4.2: (cont.) (f) forest, (g) sea, (h) cityscape

4.2 Training and Testing Methodology

Although single expert tests involve eight types of experts, in the combined tests only five types of experts are utilized. This is necessary, since using all types of experts would result in a dominance of K-NN family type experts. Therefore, in addition to the SVM, Nearest Mean and Bayesian Gaussian Plug-in type experts, only 1-NN and 5-NN type experts are used leaving out 3-NN, 7-NN and 9-NN.

During the tests conducted, for each class, 100 in-class and 100 informative out-of-class samples are used. *5-fold cross validation* is utilized in order to prevent the dependence of the performances on the data sets [39]. This is achieved by performing five tests for each class by each time taking 20 distinct samples from each of the in-class and out-of-class data sets, and training the experts using the remaining data consisting of 80 in-class and 80 out-of-class training samples. The results of these five tests are then averaged to obtain a reliable result.

Eight classes that are subjects of the tests are *football*, *indoor (outdoor)*, *crowd*, *sunset-sunrise*, *sky*, *forest*, *sea* and *cityscape*. As already mentioned in the previous section, first three of these classes are defined by a single low-level feature, and therefore only Single Feature Combination (SFC) is available as a decision mechanism. Other five are defined with multiple features and therefore the advanced decision mechanisms (MFDC and MFCC) are also applicable.

4.3 Simulation Results

Throughout the experiments, the performances of single experts and combined experts on the eight selected semantic classes are tested. Moreover, in order to

provide a good basis of comparison, for each class, the result of an "optimal combination formula", which is obtained by combining experts with the best results, is also included. Obviously, such a case is not practical, since the combination formula is non-normative and should be determined in a case-by-case basis for each class. However, for specific visual classes optimal formula can still be obtained in order to classify a large database based on that formulation.

The performances of the techniques are presented separately for each class in the following tables. The results for the first three classes, namely *football*, *indoor (outdoor)* and *crowd*, each depending on a single feature, are given in Table 4.2, Table 4.3 and Table 4.4 respectively. Other five classes, which are *sunset-sunrise*, *sky*, *forest*, *sea* and *cityscape* are presented in Table 4.5, Table 4.6, Table 4.7, Table 4.8 and Table 4.9 respectively. Three important statistics given in the tables are accuracy, precision, and recall. Accuracy is the overall classification performance which can be represented as the ratio of the number of correctly classified samples to the total number of samples. Precision can be briefly defined as the ratio of the number of samples correctly classified as in-class to the number of samples classified as in-class. Recall, on the other hand, is the ratio of the number of samples correctly classified as in-class to the number of in-class samples. Although the performance comparisons are made firstly according to accuracy, precision and recall results are also included in the tables. This is due to the fact that they convey information about different properties of the techniques, which is hidden in accuracy. Comments on the precision values are made wherever necessary because according to the preliminary experiments, precision gives important clues about the performance of the system on

Table 4.2: Football classification results

FOOTBALL		Max Single	SFC					
			Prd	Sum	Max	Min	Med	MV
Test Set 1	Accuracy	87.5	82.5	87.5	85.0	85.0	87.5	92.5
	Precision	82.6	100.0	100.0	100.0	100.0	100.0	90.5
	Recall	95.0	65.0	75.0	70.0	70.0	75.0	95.0
Test Set 2	Accuracy	95.0	92.5	97.5	92.5	92.5	97.5	97.5
	Precision	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Recall	90.0	85.0	95.0	85.0	85.0	95.0	95.0
Test Set 3	Accuracy	95.0	95.0	95.0	95.0	95.0	97.5	95.0
	Precision	95.0	100.0	100.0	100.0	100.0	100.0	95.0
	Recall	95.0	90.0	90.0	90.0	90.0	95.0	95.0
Test Set 4	Accuracy	90.0	80.0	80.0	77.5	77.5	82.5	85.0
	Precision	94.4	100.0	100.0	92.3	92.3	100.0	88.9
	Recall	85.0	60.0	60.0	60.0	60.0	65.0	80.0
Test Set 5	Accuracy	87.5	87.5	87.5	87.5	87.5	85.0	85.0
	Precision	85.7	94.1	94.1	94.1	94.1	93.8	88.9
	Recall	90.0	80.0	80.0	80.0	80.0	75.0	80.0
5-fold Cross Validation	Accuracy	91.0	87.5	89.5	87.5	87.5	90.0	91.0
	Precision	91.6	98.8	98.8	97.3	97.3	98.8	92.7
	Recall	91.0	76.0	80.0	77.0	77.0	81.0	89.0

Table 4.3: Indoor-Outdoor classification results

INDOOR (OUTDOOR)		Max Single	SFC					
			Prd	Sum	Max	Min	Med	MV
Test Set 1	Accuracy	87.5	90.0	90.0	92.5	92.5	85.0	90.0
	Precision	85.7	100.0	100.0	100.0	100.0	100.0	100.0
	Recall	90.0	80.0	80.0	85.0	85.0	70.0	80.0
Test Set 2	Accuracy	77.5	72.5	72.5	70.0	70.0	72.5	75.0
	Precision	73.9	80.0	80.0	75.0	75.0	80.0	81.3
	Recall	85.0	60.0	60.0	60.0	60.0	60.0	65.0
Test Set 3	Accuracy	77.5	85.0	80.0	82.5	82.5	77.5	82.5
	Precision	76.2	88.9	87.5	84.2	84.2	86.7	88.2
	Recall	80.0	80.0	70.0	80.0	80.0	65.0	75.0
Test Set 4	Accuracy	92.5	92.5	92.5	95.0	95.0	90.0	95.0
	Precision	94.7	100.0	100.0	100.0	100.0	100.0	100.0
	Recall	90.0	85.0	85.0	90.0	90.0	80.0	90.0
Test Set 5	Accuracy	80.0	80.0	80.0	77.5	77.5	80.0	77.5
	Precision	75.0	87.5	87.5	82.4	82.4	87.5	82.4
	Recall	90.0	70.0	70.0	70.0	70.0	70.0	70.0
5-fold Cross Validation	Accuracy	83.0	84.0	83.0	83.5	83.5	81.0	84.0
	Precision	81.1	91.3	91.0	88.3	88.3	90.8	90.4
	Recall	87.0	75.0	73.0	77.0	77.0	69.0	76.0

Table 4.4: Crowd classification results

CROWD		Max Single	SFC					
			Prd	Sum	Max	Min	Med	MV
Test Set 1	Accuracy	85.0	77.5	80.0	80.0	80.0	75.0	82.5
	Precision	88.9	76.2	80.0	77.3	77.3	75.0	84.2
	Recall	80.0	80.0	80.0	85.0	85.0	75.0	80.0
Test Set 2	Accuracy	90.0	87.5	92.5	87.5	87.5	92.5	87.5
	Precision	90.0	82.6	90.5	82.6	82.6	90.5	85.7
	Recall	90.0	95.0	95.0	95.0	95.0	95.0	90.0
Test Set 3	Accuracy	77.5	77.5	82.5	80.0	80.0	80.0	80.0
	Precision	82.4	76.2	88.2	77.3	77.3	87.5	83.3
	Recall	70.0	80.0	75.0	85.0	85.0	70.0	75.0
Test Set 4	Accuracy	75.0	65.0	72.5	67.5	67.5	72.5	72.5
	Precision	81.3	65.0	76.5	66.7	66.7	80.0	76.5
	Recall	65.0	65.0	65.0	70.0	70.0	60.0	65.0
Test Set 5	Accuracy	70.0	70.0	77.5	70.0	70.0	75.0	70.0
	Precision	75.0	62.5	73.9	62.5	62.5	75.0	68.2
	Recall	60.0	100.0	85.0	100.0	100.0	75.0	75.0
5-fold Cross Validation	Accuracy	79.5	75.5	81.0	77.0	77.0	79.0	78.5
	Precision	83.5	72.5	81.8	73.3	73.3	81.6	79.6
	Recall	73.0	84.0	80.0	87.0	87.0	75.0	77.0

Table 4.5: Sunset-Sunrise classification results

SUNSET-SUNRISE		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Test Set 1	Accuracy	95.0	95.0	90.0	90.0	92.5	92.5	90.0	90.0	90.0	92.5	92.5	92.5	92.5	92.5	97.5	PRD (CSD-1NN + CSD-NM)
	Precision	100.0	95.0	100.0	94.4	100.0	100.0	94.4	94.4	100.0	100.0	100.0	100.0	94.7	94.7	100.0	
	Recall	90.0	95.0	80.0	85.0	85.0	85.0	85.0	85.0	80.0	85.0	85.0	85.0	90.0	90.0	95.0	
Test Set 2	Accuracy	95.0	95.0	92.5	95.0	92.5	92.5	95.0	95.0	92.5	92.5	90.0	92.5	95.0	95.0	92.5	
	Precision	95.0	95.0	94.7	95.0	94.7	94.7	95.0	95.0	94.7	94.7	90.0	94.7	95.0	95.0	94.7	
	Recall	95.0	95.0	90.0	95.0	90.0	90.0	95.0	95.0	90.0	90.0	90.0	90.0	95.0	95.0	90.0	
Test Set 3	Accuracy	90.0	87.5	92.5	87.5	90.0	90.0	87.5	87.5	92.5	90.0	80.0	90.0	87.5	87.5	90.0	
	Precision	86.4	80.0	87.0	82.6	86.4	86.4	82.6	82.6	87.0	86.4	73.1	86.4	82.6	82.6	86.4	
	Recall	95.0	100.0	100.0	95.0	95.0	95.0	95.0	95.0	100.0	95.0	95.0	95.0	95.0	95.0	95.0	
Test Set 4	Accuracy	90.0	92.5	92.5	90.0	92.5	92.5	90.0	90.0	92.5	90.0	72.5	85.0	90.0	90.0	92.5	
	Precision	86.4	90.5	94.7	94.4	94.7	94.7	94.4	94.4	94.7	94.4	68.0	85.0	94.4	94.4	90.5	
	Recall	95.0	95.0	90.0	85.0	90.0	90.0	85.0	85.0	90.0	85.0	85.0	85.0	85.0	85.0	95.0	
Test Set 5	Accuracy	92.5	90.0	95.0	87.5	92.5	92.5	87.5	87.5	95.0	90.0	87.5	95.0	90.0	90.0	95.0	
	Precision	87.0	83.3	90.9	89.5	87.0	87.0	89.5	89.5	90.9	90.0	80.0	90.9	86.4	86.4	90.9	
	Recall	100.0	100.0	100.0	85.0	100.0	100.0	85.0	85.0	100.0	90.0	100.0	100.0	95.0	95.0	100.0	
5-fold Cross Validation	Accuracy	92.5	92.0	92.5	90.0	92.0	92.0	90.0	90.0	92.5	91.0	84.5	91.0	91.0	91.0	93.5	
	Precision	90.9	88.8	93.5	91.2	92.6	92.6	91.2	91.2	93.5	93.1	82.2	91.4	90.6	90.6	92.5	
	Recall	95.0	97.0	92.0	89.0	92.0	92.0	89.0	89.0	92.0	89.0	91.0	91.0	92.0	92.0	95.0	

Table 4.6: Sky classification results

SKY		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Test Set 1	Accuracy	80.0	85.0	92.5	90.0	92.5	92.5	90.0	90.0	92.5	92.5	75.0	82.5	90.0	90.0	92.5	SUM (CSD-SVM + CSD-1NN + HTD-5NN)
	Precision	71.4	76.9	87.0	83.3	87.0	87.0	83.3	83.3	87.0	87.0	75.0	88.2	83.3	83.3	87.0	
	Recall	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	75.0	100.0	100.0	100.0	
Test Set 2	Accuracy	97.5	92.5	97.5	100.0	95.0	95.0	100.0	100.0	97.5	100.0	95.0	95.0	100.0	100.0	97.5	
	Precision	95.2	94.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
	Recall	100.0	90.0	95.0	100.0	90.0	90.0	100.0	100.0	95.0	100.0	90.0	90.0	100.0	100.0	95.0	
Test Set 3	Accuracy	97.5	97.5	97.5	100.0	95.0	97.5	100.0	100.0	97.5	100.0	82.5	85.0	97.5	97.5	97.5	
	Precision	95.2	100.0	95.2	100.0	90.9	95.2	100.0	100.0	95.2	100.0	93.3	100.0	95.2	95.2	95.2	
	Recall	100.0	95.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	70.0	100.0	100.0	100.0	
Test Set 4	Accuracy	100.0	97.5	97.5	95.0	92.5	95.0	95.0	95.0	97.5	95.0	87.5	95.0	97.5	97.5	100.0	
	Precision	100.0	100.0	95.2	95.0	94.7	95.0	95.0	95.0	95.2	95.0	85.7	100.0	95.2	95.2	100.0	
	Recall	100.0	95.0	100.0	95.0	90.0	95.0	95.0	95.0	100.0	95.0	90.0	90.0	100.0	100.0	100.0	
Test Set 5	Accuracy	90.0	90.0	95.0	97.5	95.0	95.0	97.5	97.5	95.0	97.5	75.0	85.0	92.5	92.5	92.5	
	Precision	83.3	90.0	95.0	95.2	100.0	100.0	95.2	95.2	95.0	95.2	77.8	100.0	87.0	87.0	90.5	
	Recall	100.0	90.0	95.0	100.0	90.0	90.0	100.0	100.0	95.0	100.0	70.0	70.0	100.0	100.0	95.0	
5-fold Cross Validation	Accuracy	93.0	92.5	96.0	96.5	94.0	95.0	96.5	96.5	96.0	97.0	83.0	88.5	95.5	95.5	96.0	
	Precision	89.0	92.3	94.5	94.7	94.5	95.4	94.7	94.7	94.5	95.4	86.4	97.6	92.2	92.2	94.5	
	Recall	100.0	94.0	98.0	99.0	94.0	95.0	99.0	99.0	98.0	99.0	79.0	79.0	100.0	100.0	98.0	

Table 4.7: Forest classification results

FOREST		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Test Set 1	Accuracy	77.5	82.5	90.0	85.0	85.0	85.0	85.0	85.0	90.0	82.5	70.0	85.0	80.0	80.0	85.0	MAX (CSD-5NN + EHD-1NN + HTD-SVM)
	Precision	76.2	84.2	94.4	85.0	88.9	88.9	85.0	85.0	94.4	81.0	68.2	93.8	77.3	77.3	81.8	
	Recall	80.0	80.0	85.0	85.0	80.0	80.0	85.0	85.0	85.0	85.0	75.0	75.0	85.0	85.0	90.0	
Test Set 2	Accuracy	90.0	90.0	92.5	95.0	92.5	92.5	95.0	95.0	92.5	92.5	90.0	92.5	92.5	92.5	92.5	
	Precision	94.4	94.4	94.7	95.0	94.7	94.7	95.0	95.0	94.7	94.7	90.0	94.7	94.7	94.7	94.7	
	Recall	85.0	85.0	90.0	95.0	90.0	90.0	95.0	95.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	
Test Set 3	Accuracy	70.0	82.5	87.5	85.0	87.5	87.5	82.5	82.5	87.5	82.5	77.5	82.5	82.5	82.5	92.5	
	Precision	72.2	88.2	89.5	85.0	89.5	89.5	84.2	84.2	89.5	88.2	78.9	88.2	84.2	84.2	94.7	
	Recall	65.0	75.0	85.0	85.0	85.0	85.0	80.0	80.0	85.0	75.0	75.0	75.0	80.0	80.0	90.0	
Test Set 4	Accuracy	77.5	72.5	80.0	80.0	77.5	77.5	80.0	80.0	80.0	77.5	72.5	77.5	77.5	77.5	77.5	
	Precision	72.0	68.0	71.4	71.4	70.4	70.4	71.4	71.4	71.4	70.4	64.5	69.0	70.4	70.4	73.9	
	Recall	90.0	85.0	100.0	100.0	95.0	95.0	100.0	100.0	100.0	95.0	100.0	100.0	95.0	95.0	85.0	
Test Set 5	Accuracy	80.0	82.5	82.5	85.0	82.5	82.5	85.0	85.0	82.5	87.5	80.0	80.0	82.5	82.5	77.5	
	Precision	77.3	88.2	78.3	85.0	78.3	78.3	85.0	85.0	78.3	85.7	75.0	75.0	78.3	78.3	73.9	
	Recall	85.0	75.0	90.0	85.0	90.0	90.0	85.0	85.0	90.0	90.0	90.0	90.0	90.0	90.0	85.0	
5-fold Cross Validation	Accuracy	79.0	82.0	86.5	86.0	85.0	85.0	85.5	85.5	86.5	84.5	78.0	83.5	83.0	83.0	85.0	
	Precision	78.4	84.6	85.7	84.3	84.3	84.3	84.1	84.1	85.7	84.0	75.3	84.1	81.0	81.0	83.8	
	Recall	81.0	80.0	90.0	90.0	88.0	88.0	89.0	89.0	90.0	87.0	86.0	86.0	88.0	88.0	88.0	

Table 4.8: Sea classification results

SEA		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Test Set 1	Accuracy	77.5	65.0	75.0	75.0	75.0	75.0	75.0	42.5	75.0	67.5	65.0	75.0	65.0	65.0	65.0	PRODUCT (CLD-MED+HTD-MAX)
	Precision	82.4	80.0	100.0	100.0	100.0	100.0	100.0	33.3	100.0	100.0	71.4	100.0	100.0	100.0	100.0	
	Recall	70.0	40.0	50.0	50.0	50.0	50.0	50.0	15.0	50.0	35.0	50.0	50.0	30.0	30.0	30.0	
Test Set 2	Accuracy	90.0	92.5	85.0	85.0	85.0	85.0	85.0	72.5	85.0	87.5	72.5	77.5	87.5	87.5	92.5	
	Precision	83.3	87.0	85.0	85.0	85.0	85.0	85.0	71.4	85.0	89.5	71.4	78.9	89.5	89.5	94.7	
	Recall	100.0	100.0	85.0	85.0	85.0	85.0	85.0	75.0	85.0	85.0	75.0	75.0	85.0	85.0	90.0	
Test Set 3	Accuracy	72.5	77.5	82.5	82.5	82.5	82.5	82.5	52.5	82.5	75.0	72.5	75.0	77.5	77.5	80.0	
	Precision	65.5	72.0	78.3	78.3	78.3	78.3	78.3	52.0	78.3	72.7	71.4	75.0	78.9	73.9	83.3	
	Recall	95.0	90.0	90.0	90.0	90.0	90.0	90.0	65.0	90.0	80.0	75.0	75.0	75.0	85.0	75.0	
Test Set 4	Accuracy	72.5	85.0	90.0	90.0	90.0	90.0	90.0	55.0	90.0	95.0	77.5	82.5	85.0	80.0	90.0	
	Precision	64.5	79.2	86.4	86.4	86.4	86.4	86.4	54.2	86.4	90.9	76.2	84.2	79.2	73.1	94.4	
	Recall	100.0	95.0	95.0	95.0	95.0	95.0	95.0	65.0	95.0	100.0	80.0	80.0	95.0	95.0	85.0	
Test Set 5	Accuracy	90.0	95.0	97.5	97.5	97.5	97.5	97.5	77.5	97.5	97.5	82.5	85.0	97.5	95.0	100.0	
	Precision	83.3	90.9	95.2	95.2	95.2	95.2	95.2	69.0	95.2	95.2	76.0	79.2	95.2	90.9	100.0	
	Recall	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.0	95.0	100.0	100.0	100.0	
5-fold Cross Validation	Accuracy	80.5	83.0	86.0	86.0	86.0	86.0	86.0	60.0	86.0	84.5	74.0	79.0	82.5	81.0	85.5	
	Precision	75.8	81.8	89.0	89.0	89.0	89.0	89.0	56.0	89.0	89.7	73.3	83.5	88.6	85.5	94.5	
	Recall	93.0	85.0	84.0	84.0	84.0	84.0	84.0	64.0	84.0	80.0	75.0	75.0	77.0	79.0	76.0	

Table 4.9: Cityscape classification results

CITYSCAPE		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Test Set 1	Accuracy	90.0	90.0	82.5	90.0	82.5	82.5	90.0	90.0	82.5	87.5	80.0	85.0	85.0	85.0	77.5	MED (CSD-SVM+EHD-BAYES+HTD-BAYES)
	Precision	86.4	90.0	76.0	86.4	76.0	76.0	86.4	86.4	76.0	85.7	73.1	79.2	85.0	85.0	72.0	
	Recall	95.0	90.0	95.0	95.0	95.0	95.0	95.0	95.0	95.0	90.0	95.0	95.0	85.0	85.0	90.0	
Test Set 2	Accuracy	82.5	80.0	92.5	90.0	92.5	92.5	90.0	90.0	92.5	90.0	77.5	85.0	87.5	87.5	87.5	
	Precision	84.2	83.3	90.5	86.4	90.5	90.5	86.4	86.4	90.5	86.4	76.2	88.9	85.7	85.7	89.5	
	Recall	80.0	75.0	95.0	95.0	95.0	95.0	95.0	95.0	95.0	95.0	80.0	80.0	90.0	90.0	85.0	
Test Set 3	Accuracy	75.0	80.0	85.0	90.0	80.0	80.0	90.0	90.0	85.0	77.5	80.0	82.5	75.0	75.0	95.0	
	Precision	77.8	77.3	85.0	90.0	80.0	80.0	90.0	90.0	85.0	76.2	80.0	84.2	70.8	70.8	95.0	
	Recall	70.0	85.0	85.0	90.0	80.0	80.0	90.0	90.0	85.0	80.0	80.0	80.0	85.0	85.0	95.0	
Test Set 4	Accuracy	87.5	85.0	87.5	87.5	77.5	77.5	90.0	90.0	87.5	90.0	62.5	62.5	85.0	87.5	87.5	
	Precision	89.5	85.0	94.1	89.5	92.3	92.3	90.0	90.0	94.1	90.0	85.7	85.7	79.2	82.6	100.0	
	Recall	85.0	85.0	80.0	85.0	60.0	60.0	90.0	90.0	80.0	90.0	30.0	30.0	95.0	95.0	75.0	
Test Set 5	Accuracy	75.0	72.5	77.5	75.0	77.5	77.5	75.0	75.0	77.5	72.5	55.0	70.0	72.5	72.5	80.0	
	Precision	75.0	73.7	78.9	77.8	78.9	78.9	77.8	77.8	78.9	76.5	55.6	83.3	73.7	73.7	83.3	
	Recall	75.0	70.0	75.0	70.0	75.0	75.0	70.0	70.0	75.0	65.0	50.0	50.0	70.0	70.0	75.0	
5-fold Cross Validation	Accuracy	82.0	81.5	85.0	86.5	82.0	82.0	87.0	87.0	85.0	83.5	71.0	77.0	81.0	81.5	85.5	
	Precision	82.6	81.9	84.9	86.0	83.5	83.5	86.1	86.1	84.9	82.9	74.1	84.3	78.9	79.6	88.0	
	Recall	81.0	81.0	86.0	87.0	81.0	81.0	88.0	88.0	86.0	84.0	67.0	67.0	85.0	85.0	84.0	

large-scale databases. Also the variance of the accuracy of a mechanism among the five test sets gives important information about the stability and therefore analyzed along with the accuracy and precision values.

For the classes, where only one descriptive feature is used (i.e. *football*, *indoor* and *crowd*), it is seen that SFC leads with at least one rule except for the *football* case. However, improvements are not significant and also performance depends on the choice of the best combination strategy for each of these classes. For *football*, the majority vote rule gives the same result (91%) with the best expert, which is a 1-NN. *Indoor* class is classified slightly better than the best expert (83%) by product and majority vote combinations (84%). In *crowd* classification, sum rule reached 81% and beat 9-NN type expert, whose performance was the highest among single experts with 79.5%. In all of the above cases, an interesting relation between precision and stability (accuracy variance) is observed. In all of the three cases above, whenever precision is higher (*football* and *indoor*), stability is lower and vice versa (*crowd*).

In contrast with the above cases, significant improvements are observed in the cases, where the proposed advanced decision mechanisms are applicable. MFDC and MFCC outperform the best single expert and the best SFC for nearly all cases. The

only case in which advanced decision mechanisms do not yield better results than the best single expert is *sunset-sunrise* classification.

MFDC though being successful against single experts, could not beat the "optimal decision formula" in most of the cases. However, the "optimal combination formula" gives results that are inferior when compared with MFCC

for many classes. MFCC improves the performance of classifications, especially when its second stage combination rule is fixed to median, while SFCs in the previous stage are obtained by the product rule. This should be due to the fact that these two rules have properties, which compensate the weaknesses of each other. Product rule, though known to have many favorable properties, is a "severe" rule, since a single expert can inhibit the positive decision of all others by outputting close to zero probability [43]. Median rule, however, can be viewed as a robust average of all experts and is therefore more resilient to this kind of situation. This leads us to the observation that combining the product rule and the median rule is an effective method of increasing the modeling performance. This observation on MFCC is also supported by a performance improvement of 3.5% for sky, 6.5% for forest, 5.5% for sea and 5.0% for cityscape classification using MFCC, when compared against the best single expert. MFCC also achieves a performance improvement of at least 1-2% over even the manually selected "optimal combination formula" in these cases.

Another important fact about the performances achieved in classification of these classes using advanced decision mechanisms is the increase in stability and precision they provide. In the application of these methods to large databases with high variation compared with data sets used in experiments, usually recall values are sustained, however, precision values drop severely. Therefore methods achieving higher precision and stability are considered more as more successful, when the accuracy values are equal. In the experiments for the five semantic classes, MFCC combination has the highest accuracy. In the case of *sunset-sunrise* classification, where MFCC is beaten in terms of accuracy, it can be seen

that it has both higher precision and higher stability, meaning that the success of the "optimal combination formula" depends highly on the data used in experiments. As far as it can be seen from the results of the experiments, a general rule is that in the cases where accuracy difference is small and MFCC is better (*sky, sea, cityscape*), if precision is higher then the stability is lower and vice versa. Although the same result is expected for the cases where "optimal combination combination" has slightly better accuracy, in the only example of this (*sunset-sunrise*), MFCC is favored by both precision and stability. Another example that strengthens the success of MFCC is the *forest* classification. In this case, where MFCC has the highest accuracy improvement over the "optimal combination formula", it has also higher precision and stability. Typical classification results can also be observed at the website of the ongoing MPEG-7 compliant multimedia management system, BilVMS (http://vms.bilten.metu.edu.tr/).

CHAPTER 5

CONCLUSIONS

5.1 Summary of Thesis

Information is valuable as long as it can be retrieved. One of the greatest sources of information in today's world is obviously digital multimedia. The growing amount of digital multimedia data brings the problem of appropriate indexing and retrieval of large collections. Among many different approaches for fulfilling the task, classification based automatic indexing which requires the least user interaction is chosen. In this selection, indexes are limited to the generic semantic visual classes, in order to construct a basis for more complicated systems which might be the subject of the future research. These generic classes can be inferred from the entire image features rather than the objects inside images and therefore segmentation, which is still an unsolved problem, is not compulsory.

In the system implemented during this research, the problem of reaching the semantic information inside images that is closer to user needs is addressed. For this purpose, many methods for representing images are analyzed. As a result of the literature survey, MPEG-7, which is a newly emerging standard that aimed to combine many of the previous successful work inside its body, is selected as

the tool for representing images with low-level features. Two color and two texture descriptors that are defined by the standard are used as the low-level features. Experience gained as a result of the experimentations using this new standard that will probably become widespread in the near future is one of the greatest benefits in this research. It is not very optimistic to assume that in the future MPEG-7 descriptors, which are already extracted will be transmitted with any kind of multimedia data.

Classifying data into semantic classes is simply a pattern recognition problem. Success of pattern recognition systems is highly correlated with the performance of the underlying classifier structures used to model the concepts. In this research, four different classifiers structures are implemented. Along with the three common classifiers, a new classifier called SVM, which has been introduced recently and has become very popular because of its generalization performance, is utilized. Methods in the literature for normalizing the outputs of each of these classifiers are analyzed and some modifications, which force the confidence for positive results, are proposed. With the incorporation of these methods, calibrated probability outputs of these classifiers become appropriate for combined usage.

Reaching semantic information from low-level features is a challenging problem, which has attracted great attention from the research community. Most of the time, to define a semantic entity, training a single type of classifier with a single low-level feature is not enough. It is required to use multiple features for training experts with different classifier structures and combine the outputs of these experts to fit a model to the distribution of the members of these classes. For this reason, the most common and reliable combination rules, namely, product,

sum, max, min, median and majority vote, are utilized as the basic expert combination strategies. As a contribution to the existing literature, two advanced decision mechanisms, namely, Multiple Feature Direct Combination (MFDC) and Multiple Feature Cascaded Combination (MFCC) are developed on the basic strategies (SFC is considered as the realization of the basic mechanisms without further process). These mechanisms are proposed as standard combination frameworks, whose structures are reliable in most cases in improving the single expert results. The reason for a standard structure is the feasibility problem that is seen in most of the systems using specific weights and thresholds depending on the situation. This kind of a system would be impractical because of the need for the case-by-case selection of the best experts and their weights (like in the optimal combination rule).

5.2 Discussion on Simulation Results

The extensive experiments, where many different cases differentiated by the combination rules, features, classes and classifier structures analyzed, are performed using the proposed system. The results of the experiments are quite interesting and informative in many ways.

First of all, the methods that are incorporated for the experts to give probability outputs make the experts much more robust than their original versions. Confidence measures affecting the probability outputs directly seem to be contributing to the performances of all types of classifier structures used. However, in this success the role of the selected MPEG-7 low-level features should not be overlooked. If the utilized features were not discriminative enough,

advances in the classifier structures would not have affected the performances significantly.

Another important point that should be mentioned is the inferiority of SFC against single experts in nearly half of the cases where multiple feature combinations are available. This is due to the increase on the density of the experts trained on a single feature. In those cases, the combined performances deteriorate, since the confidence constraints utilized in the experts force them to be more conservative, while the provided single feature lacks the ability to define a semantic concept successfully. In the cases, where only SFC is available (football, indoor (outdoor), crowd), since these features are able to describe the characteristics of the semantic concept quite well, the complementary natures of the classifiers provided performance improvements.

Among two proposed advanced decision mechanisms, especially MFCC achieved significant improvements consistently, even in the cases where single experts have already had very successful high accuracies. The main reason for this improvement is the reliability and stability the combination enjoys, since experts that are successful at modeling different parts of the class distribution, are combined to complement each other. On the other hand, MFDC performs inconsistently and its performance seems to depend mainly on the shape of the distribution. As can be inferred from the tabulated results in the tables, though it has achieved nearly the same performance with MFCC in some cases, it can be surpassed by even the single expert with the best performance among the others. For MFCC, it is observed that classification performance significantly improves, when the correct combination rules are selected at each stage. For instance,

combining the product rule results of the first stage by using median rule at the second stage is found out to be quite successful in all cases, beating even the optimal combination formula. This observation can be explained by the complementary natures of the combination rules, which compensates each other's weaknesses.

As a result, the system that is implemented as an outcome of this research enabled the analysis of the performance of the combination of different techniques and different features. The performance improvements achieved by these combinations have been proven to be successful. The results of this research will be used as the starting point of the future work regarding the content based retrieval. In this future research, incorporation of segmentation using a user-friendly interactive method is planned. This is inevitable for reaching more specific semantic concepts, many of which are defined in terms of objects inside an image. Also as a new dimension of image representation, MPEG-7 shape descriptors are planned to be used in this scheme. Another important concept in content based retrieval has always been the relevance feedback. In the future system the goal will be to incorporate all the information that can be obtained from the user and therefore user will be able to guide the system during a continuous training phase.

REFERENCES

[1] D.A. Forsyth, "Benchmarks for storage and retrieval in multimedia databases," Proceedings of SPIE, Vol. 4676, pp. 240-247, 2002.

[2] Information Society Technologies Programme Report, State of the art in content-based analysis, indexing and retrieval, 2002.

[3] E. C. Yiu, "Image Classification Using Color Cues and Texture Orientation," M.S. Thesis, Massachusetts Institute of Technology, May 1996.

[4] A. Vailaya, M. Figueiredo, A. K. Jain, and H. Zhang, "Image Classification for Content-Based Indexing," IEEE Transactions on Image Processing, Vol.10, No.1, Jan 2001.

[5] S. Kim, W. Woo, and Y. Ho, "Image Retrieval using Multi-Scale Color Clustering," IEEE Proc. ICIP 2001, vol. I, pp. 666-669, Oct. 7-10, 2001.

[6] N. Serrano, A. Savakis, and J. Luo, "A Computationally Efficient Approach to Indoor / Outdoor Scene Classification,", IEEE Proc. ICPR 2002, Volume 4, pp. 40146, 2002.

[7] T. Syedamahmood and D. Petkovic, "On Describing Color and Shape Information in Images," Signal Processing-Image Communication 16, No. 1-2, pp. 15-31, 2000.

[8] J. R. Smith and S. F. Chang, "VisualSEEk: a fully automated content-based image query system," Proceedings of ACM Multimedia, pp. 87-98, Nov 1996.

[9] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System," IEEE Computer, Volume:28, Issue:9, pp. 23-32, September 1995.

[10] M. Szummer and R. W. Picard, “Indoor-Outdoor Image Classification,” Proceedings of IEEE Intl. Workshop on Content-Based Access of Image and Video Database, pp. 42-51, Jan. 1998.

[11] J. R. Ohm, F. Bunjamin, W. Liebsch, B. Makai, K. Mueller, A. Smolic, and D. Zier, “A Set of Visual Feature Descriptors and Their Combination in a Low-Level Description Scheme,” Signal Processing: Image Communication, July 2000.

[12] Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang, “A Bayesian Framework for Semantic Classification of Outdoor Vacation Images,” Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VII, vol. 3656, pp. 415-426, San Jose, CA, January 1999.

[13] F. Mokhtarian, A. K. Mackworth, “A theory of multiscale,curvature-based shape representation for planar curves,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No: 8, August 1992.

[14] D. Coquin, P. Bolon, and B. Ionescu, “Dissimilarity Measures in Color Spaces,” IEEE Proc. ICPR 2002, Volume: 1 , pp. 612-615, August 2002.

[15] wwwqbic.almaden.ibm.com

[16] www.virage.com

[17] Information Society Technologies Programme Report, SCHEMA Reference System Architecture v0.1, 2003.

[18] http://uranus.ee.auth.gr/Istorama/

[19] M. Lux, “Caliph & Emir: Semantic Annotation and Retrieval of Digital Photos,” Know-Center Graz Whitepaper, 2002.

[20] M. R. Naphade, C. Lin, J. R. Smith, B. Tseng, and S. Basu, “Learning to Annotate Video Databases,” Proceedings of SPIE, Vol. 4676, pp. 264-275, 2002.

[21] www.tecmath.de

[22] M. Soysal and A. A. Alatan, “Combining MPEG-7 Based Visual Experts For Reaching Semantics,” International Workshop VLBV03, Madrid, Spain, 18-19 September 2003.

[23] E. Esen, Ö. Önür, M. Soysal, Y. Yasaroglu, S. Tekinalp and A. A. Alatan, "A MPEG-7 Compliant Video Management System: BilVMS," European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), London, UK, April 2003.

[24] E. Esen, Y. Yaşaroğlu, Ö. Önür, M. Soysal, S. Tekinalp, and A. A. Alatan, "MPEG-7 Uyumlu Video Yönetim Sistemi," Sinyal İşleme ve İletişim Uygulamaları Kurultayı, İstanbul, 2003.

[25] http://vms.bilten.metu.edu.tr/

[26] B. S. Manjunath, P. Salembier, and T. Sikora, Introduction to MPEG-7 Multimedia Content Description Interface, John Wiley & Sons Ltd., England, 2002.

[27] ISO/IEC JTC1/SC29/WG11/W3703 MPEG-7 Multimedia Content Description Interface – Part 3 Visual, October 2000.

[28] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[29] N. Christianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, UK, 2000.

[30] M. Pontil and A. Verri, "Properties of Support Vector Machines," Massachusetts Institute of Technology Artificial Intelligence Laboratory Report, A. I. Memo No. 1612, August 1997.

[31] S. Tong and E. Chang, "Support Vector Machine Active Learning for Image Retrieval," ACM Multimedia, October 2001.

[32] V. Vapnik, "SVM Method of Estimating Density, Conditional Probability, and Conditional Density," IEEE International Symposium on Circuits and Systems, Geneva, Switzerland, May 2001.

[33] P. J. Phillips, "Support Vector Machines Applied to Face Recognition," Advances in Neural Image Information Processing Systems 11, MIT Press, 1999.

[34] T. Joachims, "Making Large-Scale SVM Learning Practical," Advances in Kernel Methods – Support Vector Learning, MIT Press, 1998.

[35] http://svmlight.joachims.org/

[36] J. C. Platt, “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods,” Advances in Large Margin Classifiers, MIT Press, 1999.

[37] H. Altınçay and M. Demirekler, “Post-processing of Classifier Outputs in Multiple Classifier Systems,” Lecture Notes in Computer Science, LNCS 2364, Springer-Verlag, pp. 159-168, June 2002.

[38] J. Arlandis, J. C. Perez-Cortes, and J. Cano, “Rejection Strategies and Confidence Measures for a k-NN Classifier in an OCR Task,” IEEE Proc. ICPR 2002, vol. 1, pp. 576-579, August 2002.

[39] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, Second Edition, John Wiley & Sons, Inc., New York, 2001.

[40] I. Hsieh and K. Fan, “Multiple Classifiers for Color Flag and Trademark Image Retrieval,” IEEE Transactions on Image Processing, Vol. 10, No. 6, June 2001.

[41] S. Procter and J. Illingworth, “Combining HMM Classifiers in a Handwritten Text Recognition System,” IEEE Proc. ICIP 1998, vol. 2, pp. 934-938, 1998.

[42] A. Aykut and A. Erçil, “Örüntü Tanımada Otomatik Sınıflandırma Birleştirme,” Sinyal İşleme ve İletişim Uygulamaları Kurultayı, İstanbul, 2003.

[43] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On Combining Classifiers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, March 1998.

[44] D. M. J. Tax, R. P. W. Duin, M. van Breukelen, “Comparison between product and mean classifier combination rules,” Proceedings of the Workshop on Statistical Pattern Recognition, 1997.