DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# FUSION AND COMBINATION METHODS FOR MULTIMODAL CONTENT-BASED MEDICAL IMAGE RETRIEVAL SYSTEM

by

ALI HOSSEINZADEH VAHID

August, 2012
İZMİR

# FUSION AND COMBINATION METHODS FOR MULTIMODAL CONTENT-BASED MEDICAL IMAGE RETRIEVAL SYSTEM

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Scince in
Computer Engineering, Program**

**by**

**ALI HOSSEINZADEH VAHID**

**August, 2012
İZMİR**

## M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"FUSION AND COMBINATION METHODS FOR MULTIMODAL CONTENT BASED MEDICAL IMAGE RETRIEVAL SYSTEM"** completed by **ALI HOSSEINZADEH VAHID** under supervision of **ASST. PROF. DR. ADİL ALPKOÇAK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst.Prof.Dr. Adil ALPKOÇAK

Supervisor

Prof.Dr. Alp KUT

(Jury Member)

Asst.Prof.Dr. Damla KUNTALP

(Jury Member)

Prof.Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

## ACKNOWLEDGMENTS

This thesis could not have been written without ***Assist. Prof. Dr. Adil Alpkoçak***who not only served as my supervisor but also encouraged and challenged me throughout my academic program. He and the other members of ***Dokuz Eylül Multimedia Information Retrieval (DEMIR)*** group guided and supported me and never accepted less than my best efforts. I thank them all.

I would like to acknowledge and extend my heartfelt gratitude to my spouse,***Ramisa***. Words alone cannot express what I owe her for her encouragement. Her kindness and patience enabled me to complete this thesis.

A special thanks to my ***Mom*** and ***Dad***, the most valuable asset of my life; for all their support throughout my life.

And especially to ***God***, who made all things possible.

Ali Hosseinzadeh Vahid

# FUSION AND COMBINATION METHODS FOR MULTIMODAL CONTENT BASED MEDICAL IMAGE RETRIEVAL SYSTEM

## ABSTRACT

In this study, weinvestigate the impact of different fusion methods of modalities for performance improvement of Content-based Image Retrieval (CBIR) systems. We first evaluated the performance of low-level features to determine the suitable one. Then we provided a comparison on effect of different distance functions such as Euclidean distance and Cosine distance on multimodal content-based medical image retrieval. Then we presented an in depth investigation on different combination methods for Multimodal CBIR systems. In this way, we show how overall system performance can be improved with combination of multimodality approach and how modalities should be combined in this manner. Furthermore, we suggest a new combination approach which is based on integrating multimodal retrieval and outperforms any other fusion techniques. For evaluation, we set up a series of experiments using ImageCLEF 2011 medical image retrieval track dataset. The results show that our combination approach improves the effectiveness of whole system ever and clearly outperforms over fusion techniques for performance of multimodal CBIR systems.

**Keywords:** Information retrieval, Content-based image retrieval, Multimodal retrieval, Medical image retrieval, Fusion methods

# İÇERİK TABANLI TIBBİ GÖRÜNTÜ ERİŞME SİSTEMLERİNDE MODALİTELERİN BİRLEŞİM YÖNTEMLERİ

## ÖZ

Bu çalışma, değişik modalitelerin farklibirleştirme yöntemlerinin etkilsiniİçerik tabanlı Görüntü Erişme Sistemlerinin performansını iyileştirmesinde araştırıyor. Bu amaç için, ilk düşük seviyeli görüntü özelliklerininuygun olanını belirlemek için performanslarını değerlendirildi. Sonra farklı mesafefonksiyonlarının etkisini belirlemek için bir karşılaştırma sağlandı. Daha sonra farklı kombinasyon yöntemleri hakkında detaylı bir özgeçmiş sunuldu. Bunun için , modalitelerin birleşiminin, bir bütün olarak sistem performansını nasıl iyileştirebilmesi gösterildi. Ayrıca, bu iyileşmeği daha da artırmak için entegre birleşim yöntemi önerildi ve değerlendirilmesi için, ImageCLEF 2011 tıbbi görüntüler veri seti kullanarak bir dizi deney kuruldu.Sonuçlar bizim onerdiğimiz kombinasyonubugüne kadar önerilen tüm birleşim yöntemlerinden daha iyi çalıştığını ve Görüntü Erişme Sistemlerinin performansını daha fazla artırdığını gösterir.

**Anahtar sözcükler**: Bilgi erişim, İçerik tabanlı görüntü erişimi, Tıbbi görüntü erişimi, Birleştirme yöntemleri

## CONTENTS

**Page**

# CHAPTER ONE
# INTRODUCTION

Medical images are playing an important role to detect anatomical and functional information of the body part for diagnosis, medical research and education. Therefore the ultimate goal of content based medical image retrieval is to provide diagnostic support to physicians or radiologists by displaying relevant past cases. Medical image retrieval can also be useful as a training tool for medical students and residents in education, follow-up studies, and for research purposes. But CBIR is more challenging in medical domain due to the complex nature of images. Choice of right features, similarity measurement criteria, indexing mechanism, and query formulation technique are main factors to consider in the design of a CBIR systems (Datta, 2008) (Dimitrovski, I., Gorgevik, D., Loskovska, S., 2007). Another major problem of CBIR systems is semantic gap which is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation (Kilic, D., Alpkocak, A., 2011). Other major limitations are as follows: Huge amount of objects to search among; incomplete query specification; incomplete image description. Moreover since each feature extracted from images just characterizes certain aspect of image content, multiple features are necessarily employed to improve the retrieval performance. Meanwhile, a special feature is not equally important for different image queries since a special feature has different importance in reflecting the content of different images. Therefore some research efforts have been reported to enhance CBIR performance by taking the multi-modality fusion approaches. But traditional work on multimodal integration has largely been heuristic-based. Still today, the understanding of how fusion works and by what it is influenced is limited. This means that the major challenge in information fusion is to find adjusted techniques for associating multiple sources of information for either decision–making or information retrieval. In the other words, Fusion of results from different modalities is crucial to improve the overall retrieval performance. However, fusion techniques have some limitations since every modality mostly deals with different parts of images, and it is unable to move the relevant object into higher ranks as a whole. Our experimentations showed that, at best condition, final result set may include all relevant retrieved documents of all fused modalities. However, the documents not appearing in any of the individual modality relevant sets might appear in final results sets when each individual modality scores are summed up. This is the limit of fusion techniques in multimodal CBIR systems. Moreover, each modality may require different normalization and similarity measures. This can be easily solved by Integrated Combination Multimodal Retrieval.

In this thesis, we first present an in-depth survey on CBIR systems and their properties then we analysis different combination methods for Multimodal and then present a formal model of multimodality combination. Then, we investigate different fusion techniques for text and visual modalities in content-based medical image retrieval using ImageCLEF 2011 medical image retrieval track dataset. We also present the impact of these methods on performance improvement of fused system. Therefore, we assessed the performance of different low level feature for visual modality. Also we considered the impact of different weighting models for textual modality. Moreover, we evaluate the impact of different distance functions on performance of CBIR. In this way, we analyze why improvements can be achieved with different methods, and how modalities should be combined. Furthermore, we propose a new combination approach, which we called Integrated Combination for Multimodal Retrieval (ICMR). Experimentation showed our proposal, ICMR, outperforms other fusion techniques.

The rest of this thesis is organized as follows: The next, Section 2, presents the preliminary definitions and gives a short survey on CBIR systems. Also brief information about multimodality combination levels and methods are explained. In section 3, we formalize the multimodal fusion techniques, then we discuss about the major points of our proposed method in section 4. Details of our experimental approach in multimodal medical image retrieval and its results are presented in follow. Finally we conclude our study and draw our future roadmap of this subject in section 6.

# CHAPTER TWO
# CONTENT BASED IMAGE RETRIEVAL SYSTEMS

The main goal of Content-based Image Retrieval (CBIR) is to search similar images based on their content using a set of salient, low-level image features for indexing and similarity evaluation. It has been an active and fast advancing research area since the 1990s. During the past decades, remarkable progress has been made in both theoretical research and system development. However, there remain many challenging research problems that continue to attract researchers from multiple disciplines.

Before introducing the fundamental theory of content-based retrieval, we will take a brief look at its development. Early work on image retrieval can be traced back to a conference on Database Techniques for Pictorial Applications was held in Florence in the late 1970s (Blaser, 1979). Since then, the application potential of image database management techniques has attracted the attention of researchers (Chang, 1979) (Chang, 1980) (Chang, 1981). Early techniques were not generally based on visual features but on the textual annotation of images. In other words, images were first annotated with text and then searched using a text-based approach from traditional database management systems.

Text-based image retrieval used traditional database techniques to manage images. Through text descriptions, images could be organized by topical or semantic hierarchies to facilitate easy navigation and browsing based on standard Boolean queries. Although image retrieval only based on text information can offer much flexibility in query formulation, however, since automatically generating descriptive texts for a wide spectrum of images is not feasible, most text-based image retrieval systems require manual annotation of images. Obviously, annotating images manually is a cumbersome and expensive task for large image databases, and is often subjective, context-sensitive and incomplete. As a result, it is difficult for the traditional text-based methods to support a variety of task-dependent queries. As well as the difference in human perception when describing the images might lead to inaccuracies during the retrieval process. Comprehensive surveys of early text-based image retrieval methods can be found in (Tamura, 1984).

In the early 1990s, as a result of advances in the Internet and new digital image technologies, the volume of digital images available to users increased dramatically. The difficulties faced by text-based retrieval became more and more severe. The efficient management of the rapidly expanding visual information became an urgent problem. This need formed the driving force behind the emergence of content-based image retrieval techniques. In 1992, the National Science Foundation of the United States organized a workshop on visual information management systems (Jain, 1992) to identify new directions in image database management systems. It was widely recognized that a more efficient and intuitive way to represent and index visual information would be based on properties that are inherent in the images themselves. Researchers from the communities of computer vision, database management, human-computer interface, and information retrieval were attracted to this field. Since then, research on content-based image retrieval has developed rapidly (Dowe, 1993) (Cawkill, 1993). Since 1997, the number of research publications on the techniques of visual information extraction, organization, indexing, user query and interaction, and database management has increased enormously. Similarly, a large number of academic and commercial retrieval systems have been developed by universities, government organizations, companies, and hospitals. Comprehensive surveys of these techniques and systems can be found in (Furht, 1995) (Rui, 1999) (Smeulders, 2000)

Content-based image retrieval uses the visual contents of an image to represent and index the image. In typical content-based image retrieval systems, the visual contents of the images in the database are extracted and described by multi-dimensional feature vectors. The feature vectors of the images in the database form a feature database. To retrieve images, users provide the retrieval system with example images or sketched figures. The system then changes these examples into its internal representation of feature vectors. The similarities /distances between the feature vectors of the query example or sketch and those of the images in the database are then calculated and retrieval is performed with the aid of an indexing scheme. The indexing scheme provides an efficient way to search for the image database. Recent retrieval systems have incorporated users' relevance feedback to modify the retrieval process in order to generate perceptually and semantically more meaningful retrieval results. In this chapter, we introduce these fundamental techniques for content-based image retrieval.

## 2.1 Image Content Descriptors

Generally speaking, image content may include both visual and semantic content. Visual content can be very general or domain specific. General visual content include color, texture, shape, spatial relationship, etc. Domain specific visual content, like human faces, is application dependent and may involve domain knowledge. Semantic content is obtained either by textual annotation or by complex inference procedures based on visual content.

A good visual content descriptor should be invariant to the accidental variance introduced by the imaging process (e.g., the variation of the illuminant of the scene). However, there is a tradeoff between the invariance and the discriminative power of visual features, since a very wide class of invariance loses the ability to discriminate between essential differences. Invariant description has been largely investigated in computer vision (like object recognition), but is relatively new in image retrieval (Burkhardt, 2000).

A visual content descriptor can be either global or local. A global descriptor uses the visual features of the whole image, whereas a local descriptor uses the visual features of regions or objects to describe the image content. To obtain the local visual descriptors, an image is often divided into parts first. The simplest way of dividing an image is to use a partition, which cuts the image into tiles of equal size and shape. A simple partition does not generate perceptually meaningful regions but is a way of representing the global features of the image at a finer resolution. A better method is to divide the image into homogenous regions according to some criterion using region segmentation algorithms that have been extensively investigated in computer vision. A more complex way of dividing an image, is to undertake a complete object segmentation to obtain semantically meaningful objects (like ball, car, horse).

Currently, automatic object segmentation for broad domains of general images is unlikely to succeed.

In this section, we will introduce some widely used techniques for extracting color and texture from images.

### *2.1.1 Visual Content Descriptors*

#### *2.1.1.1 Color*

Color is one of the most widely used visual features in image retrieval (Mathias, 1998) (Stricker, 1995) (Swain, 1991). Color features are relatively robust to changes in the background colors and are independent of image size and orientation. Considerable design and experimental work, and rigorous testing, hane been performed in MPEG-7 to arrive at efficient color descriptors for similarity matching. No single generic color descriptor exists that can be used for all foreseen applications. As a result, a range of descriptors has been standardized, each suitable for achieving specific similarity-matching functionalities. In the following, first we describe different color spaces and then a brief overview of each descriptor is provided.

- *Color Spaces*: Each pixel of the image can be represented as a point in a 3D color space. There is no agreement on which is the best. However, one of the desirable characteristics of an appropriate color space for image retrieval is its *uniformity* (Mathias, 1998). *Uniformity* means that two color pairs that are equal in similarity distance in a color space are perceived as equal by viewers. In other words, the measured proximity among the colors must be directly related to the psychological similarity among them.

  *RGB space* is a widely used color space for image display. It is composed of three color components red, green, and blue. These components are called "*additive primaries*" since a color in *RGB* space is produced by adding them together. In contrast, *CMY space* is a color space primarily used for printing. The three color components are cyan, magenta, and yellow. These three components are called "subtractive primaries" since a color in CMY space is produced through light absorption. Both RGB and CMY space are device-dependent and perceptually non-uniform.

The *CIE L*a*b* and *CIE L*u*v** spaces are device independent and considered to be perceptually uniform. They consist of a luminance or lightness component ($L$) and two chromatic components $a$ and $b$ or $u$ and $v$. CIE L*a*b* is designed to deal with subtractive colorant mixtures, while CIE L*u*v* is designed to deal with additive colorant mixtures. The transformation of RGB space to CIE L*u*v* or CIE L*a*b* space can be found in (Jain, 1989).

In *HSV* (or HSL, or HSB) space is widely used in computer graphics and is a more intuitive way of describing color. The three color components are hue, saturation (lightness) and value (brightness). The hue is invariant to the changes in illumination and camera direction and hence more suited to object retrieval. RGB coordinates can be easily translated to the HSV coordinates by a simple formula (Foley, 1990)

*HMMD* is a new color space defined by MPEG and is only used in the color structure descriptor (CSD) explained below. In HMMD color space, supported in MPEG-7, The *H* has the same meaning as hue in the HSV space, and *M* and *M* are the maximum and minimum among the *R*, *G*, and *B* values, respectively. The *D* component is defined as the difference between max and min. Only three of the four components are sufficient to describe the HMMD space. This color space can be depicted using the double cone structure. In the MPEG-7 core experiments for image retrieval, it was observed that the HMMD color space is very effective and compared favorably with the HSV color space. Note that the HMMD color space is a slight twist on the HSI color space, where the D component is scaled by the intensity value.

In the following sections, we will introduce some commonly used color descriptors:

- *Color Moments*: Color moments have been successfully used in some retrieval systems (Niblack, 1993), especially when the image contains just the object. The first order (mean), the second (variance) and the third order (skewness) color moments have been proved to be efficient and effective in representing color distributions of images (Stricker, 1995). Usually the color moment performs better if it is defined by both the L*u*v* and L*a*b* color spaces as opposed to solely by the HSV space. Using the additional third-order moment improves the overall retrieval performance compared to using only the first and second order moments. However, this third-order moment sometimes makes the feature representation more sensitive to scene changes and thus may decrease the performance. Since only 9 (three moments for each of the three color components) numbers are used to represent the color content of each image, color moments are a very compact representation compared to other color features. Due to this compactness, it may also lower the discrimination power. Usually, color moments can be used as the first pass to narrow down the search space before other sophisticated color features are used for retrieval.

- *Color Histogram*: The color histogram is easy to compute and effective in characterizing both the global and local distribution of colors in an image. In addition, it is robust to translation and rotation about the view axis and changes only slowly with the scale, occlusion and viewing angle. Since any pixel in the image can be described by three components in a certain color space (for instance, red, green, and blue components in RGB space, or hue, saturation, and value in HSV space), a histogram, i.e., the distribution of the number of pixels for each quantized bin, can be defined for each component. Clearly, the more bins a color histogram contains, the more discrimination power it has. However, a histogram with a large number of bins will not only increase the computational cost, but will also be inappropriate for building efficient indexes for image databases.
- *Color Coherence Vector (CCV)*: In (Pass, 1996) a different way of incorporating spatial information into the color histogram, color coherence vectors was proposed. Each histogram bin is partitioned into two types, i.e., coherent, if it belongs to a large uniformly-colored region, or incoherent, if it does not. Due to its additional spatial information, it has been shown that

CCV provides better retrieval results than the color histogram, especially for those images which have either mostly uniform color or mostly texture regions. In addition, for both the color histogram and color coherence vector representation, the HSV color space provides better results than CIE L*u*v* and CIE L*a*b* space.

- *Color Correlogram*: The color correlogram was proposed to characterize not only the color distributions of pixels, but also the spatial correlation of pairs of colors (Huang, 1997). The first and the second dimension of the three-dimensional histogram are the colors of any pixel pair and the third dimension is their spatial distance. A color correlogram is a table indexed by color pairs, where the $k$. th entry for $(i, j)$ specifies the probability of finding a pixel of color $j$ at a distance $k$ from a pixel of color $i$ in the image. Compared to the color histogram and CCV, the color correlogram provides the best retrieval results, but is also the most computational expensive due to its high dimensionality.
- *Scalable Color Descriptor (SCD)*: One of the most basic descriptions of color features is provided by describing color distribution in images. If such a distribution is measured over an entire image, global color features can be described. The MPEG-7 generic SCD is a color histogram encoded by a Haar transform. It uses the HSV colors space uniformly quantized to 255 bins. To arrive at a compact representation the histogram bin values are nonuniformly quantized in a range from 16 bits/histogram for a rough representation of color distribution and up to 1000 bits/histogram for high-quality applications. Matching between SCD realizations can be performed by matching Haar coefficients or histogram bin values employing an L1 norm.
- *Dominant Color Descriptor*: This color descriptor aims to describe global as well as local spatial color distribution in images for high-speed retrieval and browsing. In contrast to the Color Histogram approach, this descriptor arrives at a much more compact representation at the expense of lower performance in some applications. Colors in a given region are clustered into a small number of representative colors. The descriptor consists of the representative

colors, their percentages in a region, spatial coherency of the color, and color variance.

- *Color Layout Descriptor (CLD)*: This descriptor is designed to describe spatial distribution of color in an arbitrarily-shaped region. Color distribution in each region can be described using the Dominant Color Descriptor above. The spatial distribution of color is an effective description for sketch-based retrieval, content filtering using image indexing, and visualization.
- *Color Structured Descriptor(CSD)*: The main purpose of the CSD is to express local color features in images. To this aim, a pel structuring block scans the image in a sliding window approach. With each shift of the structuring element, the number of times a particular color is contained in the structure element is counted, and a color histogram is constructed in such a way.
- *Group-of-Frames/Group-of-Pictures (GoF/GoP) Color Descriptor*: The GoF/GoP color descriptor defines a structure required for representing color features of a collection of similar frames or video frames by means of the SCD. It is useful for retrieval in image and video databases, video shot grouping, image-to-segment matching, and similar applications. It consists of average, median, and intersection histograms of groups of frames calculated based on the individual frame histograms.

#### *2.1.1.2 Texture*

The texture information of an image is a fundamental visual feature, which has been studied during the last decade to analyze images in the areas of medical imaging and satellite imaging, etc. This contains structureness, regularity, directionality and roughness of images, which are important properties of the content-based indexing of the image (Blaser, 1979). In this section, we introduce a number of texture representations, which have been used frequently and have proved to be effective in content-based image retrieval systems:

- Tamura Features : The Tamura features (Tamura, 1978.), including coarseness, contrast, directionality, line likeness, regularity, and roughness, are designed in accordance with psychological studies on the human perception of texture. The first three components of Tamura features have been used in some early well-known image retrieval systems, such as QBIC (Niblack, 1993).

- *Wold Features*: Wold decomposition (Liu, 1996)provides another approach to describing textures in terms of perceptual properties. The three Wold components, harmonic, evanescent, and non-deterministic, correspond to periodicity, directionality, and randomness of texture respectively. Periodic textures have a strong harmonic component; highly directional textures have a strong evanescent component, and less structured textures tend to have a stronger non-deterministic component.
- *Simultaneous Auto-Regressive (SAR) Model:* The SAR model is an instance of Markov random field (MRF) models, which have been very successful in texture modeling in the past decades (Kashyap, 1983). Compared with other MRF models, SAR uses fewer parameters. In the SAR model, pixel intensities are taken as random variables. To describe textures of different granularities, the multi-resolution simultaneous auto-regressive model (MRSAR) (Mao, 1992) has been proposed to enable multi-scale texture analysis. An image is represented by a multi-resolution Gaussian pyramid with low-pass filtering and sub-sampling applied at several successive levels. Either the SAR or MRSAR model may then be applied to each level of the pyramid. MRSAR has been proved (Manjunath, 1996)to have better performance than many other texture features, such as principal component analysis, Wold decomposition, and wavelet transform.
- *Gabor Filter Features*: The Gabor filter has been widely used to extract image features, especially texture features (Jain, 1991). It is optimal in terms of minimizing the joint uncertainty in space and frequency, and is often used as an orientation and scale tunable edge and line (bar) detector. There have been many approaches proposed to characterize textures of images based on Gabor filters.
- *Wavelet Transform Features*: Similar to the Gabor filtering, the wavelet transform (Daubechies, 1990)provides a multi-resolution approach to texture

analysis and classification (Chang, 1993). The computation of the wavelet transforms of a 2D signal involves recursive filtering and sub-sampling. At each level, the signal is decomposed into four frequency sub-bands, *LL*, *LH*, *HL*, and *HH*, where *L* denotes low frequency and *H* denotes high frequency. Two major types of wavelet transforms used for texture analysis are the *pyramid-structured wavelet transform* (PWT) and the *tree-structured wavelet transforms* (TWT). The PWT recursively decomposes the LL band. However, for some textures the most important information often appears in the middle frequency channels. To overcome this drawback, the TWT decomposes other bands such as LH, HL or HH when needed. After the decomposition, feature vectors can be constructed using the mean and standard deviation of the energy distribution of each sub-band at each level. For three-level decomposition, PWT results in a feature vector of $3 \times 4 \times 2$ components. For TWT, the feature will depend on how sub-bands at each level are decomposed. A fixed decomposition tree can be obtained by sequentially decomposing the LL, LH, and HL bands, and thus results in a feature vector of $52 \times 2$ components. Note that in this example, the feature obtained by PWT can be considered as a subset of the feature obtained by TWT. Furthermore, according to the comparison of different wavelet transform features (Ma, 1995), the particular choice of wavelet filter is not critical for texture analysis.

- *Homogeneous Texture Descriptor (HTD)*: HTD is composed of 62 numbers. The first two are the mean and the standard deviation of the image. The rest are the energy and the energy deviation of the Gabor filtered responses of the "channel", in the subdivision layout of the frequency domain. This design is based on the fact that response of the visual cortex is band limited and brain decomposes the spectra into bands in spatial frequency (Manjunath, 1996)
- *Edge Histogram Descriptor (EHD)*: The edge histogram descriptor represents local edge distribution in the image. It describes edges in each 'sub-image', which is obtained by dividing the image using 4x4 grids. Edges in the sub-image are classified into five types; vertical, horizontal, 45-degree, 135-degree, and non-directional. Occurrence of each type becomes a histogram

bin, producing 80 histogram bins overall. The histogram bin values are normalized by the total number of the image-blocks. The bin values are then non-linearly quantized to keep the size of the histogram as small as possible. Totally 3 bits/bin and 240 bits are needed (ISO/IEC/JTC1/SC29/WG11, 2000).

#### *2.1.1.3 Compact Composite Descriptor*

In most retrieval systems that combine two or more feature types, such as color and texture, independent vectors are used to describe each kind of information. It is possible to achieve very good retrieval scores by increasing the size of the descriptors, but this technique has several drawbacks. If the descriptor has hundreds or even thousands of bins, it may be of no practical use because the retrieval procedure is significantly delayed. Also, increasing the size of the descriptor increases the storage requirement which may have a significant penalty for databases that contain millions of images. Many presented methods limit the length of the descriptor to a small number of bins, leaving the possible factor values in decimal, non-quantized form. Here we introduce some of such new and well known set of composite descriptors .The experimental results show that the performance of the proposed descriptors is better than the performance of the similarly-sized MPEG-7 descriptors. (Chatzichristofis, 2010)

- *Color and edge directivity descriptor (CEDD)*: The CEDD includes texture information produced by the six-bin histogram of the fuzzy system that uses the five digital filters proposed by the MPEG-7 EHD. Additionally, for color information the CEDD uses the 24-bin color histogram produced by the 24-bin fuzzy-linking system. Overall, the final histogram has 144 regions. Each Image Block interacts successively with all the fuzzy systems. In the Texture Unit, the Image Block is separated into four regions called Sub Blocks. The value of each Sub Block is the mean value of the luminosity of the pixels it contains. The luminosity values are derived from a YIQ color space transformation. Each Image Block interacts with the five digital filters proposed by MPEG-7 EHD, and with the use of the pentagonal diagram it is classified in one or more texture categories. Then, in the Color Unit, every

Image Block is converted to the HSV color space. The mean values of H, S and V are calculated and become inputs to the fuzzy system that produces the fuzzy ten-bin histogram. Then, the second fuzzy system uses the mean values of S and V as well as the position number of the bin (or bins) resulting from the previous fuzzy ten-bin unit, calculates the hue of the color and produces the fuzzy 24-bin histogram. The combination of the three fuzzy systems will finally classify the Image Block. The process is repeated for all the image blocks. At the completion of the process, the histogram is normalized and quantized (Chatzichristofis, 2008).

- *Fuzzy color and texture histogram (FCTH)* : The FCTH descriptor includes the texture information produced in the eight-bin histogram of the fuzzy system that uses the high frequency bands of the Haar wavelet transform. For color information, the descriptor uses the 24-bin color histogram produced by the 24-bin fuzzy-linking system. Overall, the final histogram includes192 regions. Each Image Block interacts successively with all the fuzzy systems in the exact manner demonstrated in CEDD production. Each Image Block is transformed into the YIQ color space and transformed with the Haar Wavelet transform. The fLH, fHL and fHH values are calculated and with the use of the fuzzy system that classifies the f coeficients, this Image Block is classified in one of the eight output bins. Next, the same Image Block is transformed into the HSV color space and the mean H, S and V block values are calculated. These values become inputs to the fuzzy system that forms the ten-bin fuzzy color histogram. Then, the next fuzzy system uses the mean values of S and V as well as the position number of the bin (or bins) resulting from the previous fuzzy ten-bin unit, to calculate the hue of the color and create the fuzzy 24-bin histogram. The combined three fuzzy systems therefore classify the Image Block. The process is repeated for all the blocks of the image. At the completion of the process, the histogram is normalized and quantized (Chatzichristofis, 2008).
- *Brightness and Texture Directionality Histogram (BTDH)*: This feature is very similar to FCTH feature. The main difference from FCTH feature is using brightness instead of color histogram. This descriptor uses brightness

and texture characteristics as well as the spatial distribution of these characteristics in one compact 1D vector. The most important characteristic of the proposed descriptor is that its size adapts according to the storage capabilities of the application that is using it. This characteristic renders the descriptor appropriate for use in large medical (or gray scale) image databases. To extract the proposed descriptor, a two unit fuzzy system is used. To extract the brightness information, a fuzzy unit classifies the brightness values of the image's pixels into L_{Bright} clusters. The cluster centers are calculated using the Gustafson Kessel Fuzzy Classifier. The texture information embodied in the proposed descriptor comes from the Directionality histogram. This feature is part of the well known Tamura texture features. Fractal Scanning method through the Hilbert Curve or the Z-Grid method is used to capture the spatial distribution of brightness and texture information (Chatzichristofis, 2010)

## 2.2 Similarity Measures and Indexing Schemes

Instead of exact matching, content-based image retrieval calculates visual similarities between a query image and images in a database. Accordingly, the retrieval result is not a single image but a list of images ranked by their similarities with the query image. Many similarity measures have been developed for image retrieval based on empirical estimates of the distribution of features in recent years. Different similarity/distance measures will affect retrieval performances of an image retrieval system significantly. They are classified into three categories according to their theoretical origins (Hu, 2008). In this section, we will introduce some commonly used similarity measures.

### *2.2.1 Geometric Measures*

Geometric measures treat objects as vectors. We denote a score function $\mathrm{f_s}(\mathrm{d},\mathrm{q})$ assigns each document $\mathrm{d}$ in the document set $\mathrm{D}$ of $\mathrm{n}$ document, $\mathrm{D} = \{\mathrm{d_1}, \mathrm{d_2}, .., \mathrm{d_n}\}$, a real number which is the measure value of its similarity to a query image $q$ in query set $Q$. Some famous members of this category are:

*Minkowski Family:* Minkowski-form distance is the most widely used metric for image retrieval. If dimension of image feature vectors are independent of each other and has the same importance, the minkowski-form measures are appropriate for calculating the distance between two images. This distance is defined as:

$$f_s(d,q) = \left(\sum_i |d_i - q_i|^p\right)^{\frac{1}{p}}$$

When $\mathrm{p} = 1$, it is called the city block or Manhattan distance and defined as:

$$f_s(d,q) = \left(\sum_i |d_i - q_i|\right)$$

When $\mathrm{p} = 2$, it is called Euclidean distance and defined as:

$$f_s(d,q) = \left(\sum_i |d_i - q_i|^2\right)^{\frac{1}{2}}$$

*Cosine Function Based*: Given two vectors of attributes, the cosine similarity is represented using a dot product and magnitude as:

$$f_s(d,q) = \frac{d.q}{\|d\|\|q\|} = \frac{\sum_{i=1}^{n}(d_i \times q_i)}{\sqrt{\sum_{i=1}^{n}(d_i)^2} \times \sqrt{\sum_{i=1}^{n}(q_i)^2}}$$

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

*The Canberra distance*: $d^{CAD}$ between two vectors in an n-dimensional real vector space is given as follows:

$$f_s(d,q) = \sum_{i=1}^{n} \frac{|d_i - q_i|}{|d_i| + |q_i|}$$

### 2.2.2 *Information Theoretic Measures*

Information Theoretic Measures are derived from the Shannon's entropy theory and treat objects as probabilistic distributions. The most famous of them are:

*Kullback-Leibler (K-L) Divergence* :

$$f_s(d,q) = \sum_{i=1}^{n} d_i \, ln \frac{d_i}{q_i}$$

*Jeffrey Divergence*:

$$f_s(d,q) = \sum_{i=1}^{n} (d_i - q_i) \, ln \frac{d_i}{q_i}$$

### 2.2.3 *Statistic Measures*

Compare two objects in a distributed manner, and basically assume that the vector elements are samples.

$X^2$*Statistics* [8]:

$$f_s(d,q) = \sum_{i=1}^{n} \frac{(d_i - m_i)^2}{m_i} \quad where \quad m_i = \frac{d_i + q_i}{2}$$

## 2.3 User Interaction

For content-based image retrieval, user interaction with the retrieval system is crucial since flexible formation and modification of queries can only be obtained by involving the user in the retrieval procedure. User interfaces in image retrieval systems typically consist of a query formulation part and a result presentation part.

### *2.3.1 Query Specification*

Specifying what kind of images a user wishes to retrieve from the database can be done in many ways. Commonly used query formations are: *category browsing, query by concept, query by sketch, and query by example*. Category browsing is to browse through the database according to the category of the image. For this purpose, images in the database are classified into different categories according to their semantic or visual content. Query by concept is to retrieve images according to the conceptual description associated with each image in the database. Query by sketch and query by example is to draw a sketch or provide an example image from which images with similar visual features will be extracted from the database. The first two types of queries are related to the semantic description of images. Query by sketch allows user to draw a sketch of an image with a graphic editing tool provided either by the retrieval system or by some other software. Queries may be formed by drawing several objects with certain properties like color, texture, shape, sizes and locations. In most cases, a coarse sketch is sufficient, as the query can be refined based on retrieval results. Query by example allows the user to formulate a query by providing an example image. The system converts the example image into an internal representation of features. Images stored in the database with similar features are then searched. Query by example can be further classified into query by external image example, if the query image is not in the database, and query by internal image example, if otherwise. For query by internal image, all relationships between images can be pre-computed.

The main advantage of query by example is that the user is not required to provide an explicit description of the target, which is instead computed by the system. It is suitable for applications where the target is an image of the same object or set of objects under different viewing conditions. Most of the current systems provide this form of querying. Query by group example allows user to select multiple images. The system will then find the images that best match the common characteristics of the group of examples. In this way, a target can be defined more precisely by specifying the relevant feature variations and removing irrelevant variations in the query. In addition, group properties can be refined by adding negative examples. Many recently developed systems provide both queries by positive and negative examples.

### *2.3.2 Relevance Feedback*

Human perception of image similarity is subjective, semantic, and task-dependent. Although content-based methods provide promising directions for image retrieval, generally, the retrieval results based on the similarities of pure visual features are not necessarily perceptually and semantically meaningful. In addition, each type of visual feature tends to capture only one aspect of image property and it is usually hard for a user to specify clearly how different aspects are combined. To address these problems, interactive relevance feedback, a technique in traditional text-based information retrieval systems, was introduced. With relevance feedback, it is possible to establish the link between high-level concepts and low-level features.

Relevance feedback is a supervised active learning technique used to improve the effectiveness of information systems. The main idea is to use positive and negative examples from the user to improve system performance. For a given query, the system first retrieves a list of ranked images according to a predefined similarity metrics. Then, the user marks the retrieved images as relevant (positive examples) to the query or not relevant (negative examples). The system will refine the retrieval results based on the feedback and present a new list of images to the user. Hence, the key issue in relevance feedback is how to incorporate positive and negative examples to refine the query and/or to adjust the similarity measure.

## 2.4 Performance Evaluation

To evaluate the performance of retrieval system, two measurements, namely, recall and precision (Smeulders, 2000), are borrowed from traditional information retrieval. *Precision* is a measure that evaluates the efficiency of a system according to relevant items only, beside this; *Recall* determines the retrieval efficiency according to all relevant documents. Precision and recall is formalized as:

$$Recall = \frac{Number\ of\ relavant\ items\ retrieved}{Total\ number\ or\ retrieved\ items\ in\ the\ collection}$$

$$Precision = \frac{Number\ of\ relavant\ items\ retrieved}{Number\ of\ relevant\ items\ in\ the\ collection}$$

Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the top most results returned by the system. This measure is called *precision at n* or *P@n*.

Usually, a tradeoff must be made between these two measures since improving one will sacrifice the other. In typical retrieval systems, recall tends to increase as the number of retrieved items increases; while at the same time the precision is likely to decrease. In addition, selecting a relevant data set is much less stable due to various interpretations of the images. Further, when the number of relevant images is greater than the number of the retrieved images, recall is meaningless. As a result, precision and recall are only rough descriptions of the performance of the retrieval system. Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. By computing a precision and recall at every position in the ranked sequence of documents, one can plot a precision-recall curve, plotting precision $\mathrm{p}(\mathrm{r})$as a function of recall . *Average precision* computes the average value of $\mathrm{p}(\mathrm{r})$ over the interval from $\mathrm{r} = 0$ to $\mathrm{r} = 1$:

$$\mathrm{Avg}\, P = \int_0^1 p(r)\, dr$$

This integral is in practice replaced with a finite sum over every position in the ranked sequence of documents:

$$Avg\, P = \sum_{k=1}^{n} p(k)\, \Delta r\, (k)$$

where $\mathrm{k}$ is the rank in the sequence of retrieved documents, $\mathrm{n}$ is the number of retrieved documents, $\mathrm{p}(\mathrm{k})$ is the precision at cut-off $\mathrm{k}$in the list, and $\Delta\mathrm{r}(\mathrm{k})$ is the change in recall from items $\mathrm{k} - 1$ to $\mathrm{k}$ . This finite sum is equivalent to:

$$Avg\, P = \frac{\sum_{k=1}^{n} \big(p(k) \times rel(k)\big)}{Number\ of\ relevant\ documents}$$

where $\mathrm{rel}(\mathrm{k})$ is an indicator function equaling 1 if the item at rank $\mathrm{k}$ is a relevant document, zero otherwise. Note that the average is over all relevant documents and the relevant documents not retrieved get a precision score of zero.

*R-Precision* is another method for calculating document level averages. R-precision is the precision value of system after R documents are retrieved and R is the number of relevant documents for the topic. This method loses the impact of exact ranking of retrieved relevant documents. This measure is highly correlated to Average Precision. Also, Precision is equal to Recall at the $R$-th position.

*Mean Average Precision (MAP)* for a set of queries is the mean of the average precision scores for each query:

$$MAP = \frac{\sum_{q=1}^{Q} Avg\ P(q)}{Q}$$

where Q is the number of queries. The major difference between MAP and other evaluation measures is that MAP provides a single-figure measure of quality across recall levels. Also MAP has especially good discrimination and stability among others.

*Bpref* is designed for situations where relevance judgments are known to be far from complete. Bpref computes a preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant documents.

# CHAPTER THREE
# INFORMATION FUSION

Recently, the vast number of disparate research areas utilizes some form of information fusion in their context of theory. The fusion of multiple modalities can provide complementary information and increase the performance of the overall IR system. Bostrom and et al reviewed previous definitions of information fusion and proposed a novel definition based on strengths and weaknesses of existing definitions:

"*Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making*." (Boström, 2007)

Characteristics of multiple modalities influence the way that fusion process is carried out. Some of these properties are:

- The processing time of different types of media streams are dissimilar, which influences the fusion strategy that needs to be adopted.
- The modalities may be correlated or independent. The correlation can be perceived at different levels, such as the correlation among low-level features that are extracted from different media streams and the correlation among semantic-level decisions that are obtained based on different streams. On the other hand, the independence among the modalities is also important as it may provide additional cues in obtaining a decision. When fusing multiple modalities, this correlation and independency may equally provide valuable insight based on a particular scenario or context.
- The different modalities usually have varying confidence levels in accomplishing different tasks.
- The capturing and processing of media streams may involve certain costs, which may influence the fusion process. The cost may be incurred in units of time, money or other units of measure.

Due to these varying characteristics and the objective tasks that need to be carried out, several challenges may appear in the multimodal fusion process as stated in the following:

***Levels of fusion***: One of the earliest considerations is to decide what strategy to follow when fusing multiple modalities. The most widely used strategy is to fuse the information at the feature level, which is also known as *early fusion*. The other approach is decision level fusion or *late fusion* which fuses multiple modalities in the semantic space (Snoek, 2005). A combination of these approaches is also practiced as the hybrid fusion approach (Wu, 2006).

***How to fuse?*** There are several methods that are used in fusing different modalities. These methods are particularly suitable under different settings and are described in this section in greater detail. The discussion also includes how the fusion process utilizes the feature and decision level correlation among the modalities (Poh, 2005).

***When to fuse?*** The time when the fusion should take place is an important consideration in the multimodal fusion process. Certain characteristics of media, such as varying data capture rates and processing time of the media, poses challenges on how to synchronize the overall process of fusion. Often this has been addressed by performing the multimedia analysis tasks (such as event detection) over a timeline (Chieu, 2004). A timeline refers to a measurable span of time with information denoted at designated points. The timeline-based accomplishment of a task requires identification of designated points at which fusion of data or information should take place. Due to the asynchrony and diversity among streams and due to the fact that different analysis tasks are performed at different granularity levels in time, the identification of these designated points, i.e. when the fusion should take place, is a challenging issue.

***What to fuse?*** The different modalities used in a fusion process may provide complementary or contradictory information and therefore knowing which modalities are contributing towards accomplishing an analysis task needs to be understood. This is also related to finding the optimal number of media streams (Wu, 2004) or feature sets required to accomplish an analysis task under the specified constraints. If the most suitable subset is unavailable, can one use alternate streams without much loss of cost-effectiveness and confidence?

In this section we investigate on different levels of multimodal fusion, their characteristics, advantages and limitations.

## 3.1 Levels of Fusion

The fusion of different modalities is generally performed at two levels: *feature level* or *early fusion* and *decision level* or *late fusion*. Some researchers have also followed a hybrid approach by performing fusion at the feature as well as the decision level.

*Early fusion* is also called low-level or feature level fusion. It is an information process that integrates associates, correlates and combines uni-modal features, data and information from single or multiple sensors or sources to achieve refined estimates of parameters, characteristics, events and behaviors (Llinas J., 2004). One of the most significant downsides of this approach is that the features to be fused should be represented in the same format before fusion (Snoek C., 2005). Besides, another important difficulty of this tactic is the time synchronization between the multimodal features. In addition, the increase in the number of modalities makes it difficult to learn the cross-correlation among the heterogeneous features. Feature concatenation method is one of the simplest state synchronous early fusion methods. It used to concatenate feature vectors of all images in data collection and topics and offered them as the joint feature vector. Then similarity score between the joint vectors corresponding to the query example image and the dataset images can be calculated. Then images corresponding to top k similar joint vector can be reported as retrieved document set of fused modality per query. The major drawback of this method is that it is confronted with the curse of dimensionality as the dimension of the resulting feature space is equal to the sum of the dimensions of the subspaces. High–dimensional spaces tend to scatter the homogeneous clusters of instances belonging to the same concepts. This has to be handled using an appropriate feature weighting scheme, which is usually difficult to achieve in practice for complex multi-class problems where the majority of features are important to predict one particular class but introduce noise for all the other classes.

*Late fusion* is also called high-level or score level fusion. Here, each modality feature is, first, processed individually. The results can be scores in classification or ranks for retrieval. The resulting scores are then combined for determining the final decision. The late information fusion can be done hierarchically and on an abstract level to combine the expert's decisions but it is seen as a very rigid solution. This type of information fusion reduces the complexity of problem due to independent processing of individual modalities and improves the scalability of problem in terms of the fused modalities. On the other hand, disadvantage of the late fusion approach lies on its failure to utilize the feature level correlation among modalities. (Atrey P. K., 2010) Generally in late fusion based multimodal retrieval system, combination applied on first top k retrieved document responding to each query. The difference between such methods appear in substitution manner for value of similarity score of documents that are in retrieved document list of any combined modalities while are not in another’s one. Some approaches substitute it with zero while some other trials lookup the real similarity value of the document in the related modality and utilize it in fusion.

When combining different modalities, there are two main approaches. The relevance of a document can be measured by either its rank in the list given by an IR system or by its similarity score to the query. The score–based strategies, although more common, require a normalization among all systems in order to balance the importance of each of them, which is not the case of the rank–based strategies. In literature (Muller, 2010), there are various methods for weight normalization that we explain some of them in below:

### *3.1.1 Weight Normalization Function*

Notice that in normalization functions formula mentioned below $\mathrm{f_x}$ is similarity score of each document and $\acute{\mathrm{f}}_\mathrm{s}$ is normalized value of it.

*Min-Max Normalization*: Min-max normalization performs a linear transformation on the original data. This linear transformation, as defined in follow, produces a set of scores in the range [1: 0], where the top score is guaranteed to be 1 and the lowest score is 0.

$$\acute{f_s} = \frac{f_x - f_{min}}{f_{max} - f_{min}}$$

where $f_{max}$and $f_{min}$are the maximum and minimum score values assigned by the IR model.

*Z-Score Normalization*: In Z-score normalization, also called zero-mean, the values for an arbitrary modality are normalized based on the mean and standard deviation of similarity scores by the following formula:

$$\acute{f_s} = \frac{f_x - \mu}{\sigma}$$

where $\mu$ is the mean of scores; $\sigma$ is the standard deviation of them. $\acute{f_s}$ is negative when the raw score is below the mean and positive when above.

*Decimal Scaling*: Normalization by decimal scaling normalizes scores by moving the decimal point of value of similarity scores. The number of decimal points moved depends on the maximum absolute value of scores.

$$\acute{f_s} = \frac{f_x}{10^m}$$

where $m$ is the smallest integer such that $Max\left(\left|\acute{f_s}\right|\right) < 1$ .

Each of these methods has pros and cons. The *Min-Max* and *Z-score* methods are preferred when the matching scores of the individual modalities can be easily computed. But these methods are sensitive to outliers.

Based on an overview of the main techniques and their interdependences in (Muller, 2010), the late fusion techniques are most widely used and developed level of multimodality fusion. Therefore in next section, we describe some famous techniques of this level.

### *3.1.2 Linear Weighted Combination*

*Linear weighted combination* is one of the simplest and most widely used methods. In this method, the information obtained from different modalities is combined in a linear fashion. To combine the information, one may assign normalized weights to different modalities. Linear combination of scores, as defined in follow, was used in a large number of papers (37% of the papers dealing with information fusion in ImageCLEF) (Muller, 2010).

$$f'_{s_{mixed}}(d) = \alpha\, f'_{s_{mod1}}(d) + \beta f'_{s_{mod2}}(d)$$

where $f'_{s_{mod1}}(d)$ and $f'_{s_{mod2}}(d)$ are the normalized similarity scores of document $d$ respond to query $q$ in different modalities and $\alpha, \beta$ coefficients are their weight respectively. Linear combinations based on ranks have the advantage of not requiring a prior normalization. However, the assessment of confidence of the modalities is lost as two images having the same rank in both textual and visual modalities can have very different relevance towards the query.A particular case of the linear combination is the *CombSUM* rule where the scores of each modality are summed to obtain the final score:

$$f'_{s_{mixed}}(d) = \sum_{i=1}^{N} f'_{s_{mod_i}}(d)$$

with $N$ is the number of modalities to be combined. A variant of the *CombSUM* method is the *CombMNZ* combination rule which aims at giving more importance to the documents retrieved by several systems as follows (Shaw and Fox, 1994):

$$f'_{s_{mixed}}(d) = K.\sum_{i=1}^{N} f'_{s_{mod_i}}(d)$$

where K is equal to the number of modalities that retrieved d. CombMNZ was slightly modified by Inkpen et al (2008) for the photo retrieval task where a weight was applied to the normalized scores of each modality in order to control their respective influences.Contrary to combSUM, the combMAX and combMIN rules put all their confidence in one single modality as follows:

$$combMAX: \qquad f'_{s_{mixed}}(d) = \arg\max_{i=1}^{N}\left(f'_{s_{mod_i}}(d)\right)$$

$$combMIN: \qquad f'_{s_{mixed}}(d) = \arg\min_{i=1}^{N}\left(f'_{s_{mod_i}}(d)\right)$$

The combPROD combination rule uses the product of scores to compute final score:

$$f'_{s_{mixed}}(d) = \prod_{i=1}^{N} f'_{s_{mod_i}}(d)$$

From another point of view (Sanderson, 2004), information fusion is classified into three main categories: *pre-mapping fusion, midst-mapping fusion* and *post-mapping fusion*. In pre-mapping fusion, information is combined before any use of classifiers or experts; in midst-mapping fusion, information is combined during mapping from sensor data/feature space into opinion/decision space, while in post mapping fusion, information is combined after mapping from sensor data/feature space into opinion/decision space.

Another categorization of fusion method presented in (Atrey P. K., 2010) and divided the fusion methods into the following three categories: rule-based methods, classification-based methods, and estimation-based methods. The rule-based fusion methods include a variety of statistical rules of combining multimodal information and their performance generally is related to quality of temporal alignment between different modalities. Classification-based fusion methods include a range of classification techniques that have been used to classify the multimodal observation into one of the pre-defined classes. The estimation-based fusion methods have been primarily used to better estimate and predict the fused observations of the state of a moving object over a period based on multimodal data. These methods are suitable for object localization and tracking tasks.

## 3.2 Limitation of Fusion

CBIR systems have become mature enough to extract semantic information that is complementary to textual information, thus allowing enhancement of the quality of retrieval both in terms of precision and recall. Early fusion enables a comprehensive overview of the multi–modal information by combining modalities inside the IR system and offers potentially high flexibility for promoting relevant modalities in the context of a particular query. Unfortunately, it is difficult to put into practice because it relies on large and heterogeneous feature spaces that become less distinctive, due to what is called the curse of dimensionality. Moreover, combining binary and categorical variable that are textual attributes with continuous and correlated visual features is not trivial and negative interactions among features can occur. Late fusion techniques are by far the most frequently utilized with more than 60% of the imageCLEF papers dealing with textual and information fusion. This is not surprising as late fusion allows for a straightforward combination of any system a given threshold. However it was observed that combining textual and visual information is not devoid of risks and can degrade the retrieval performance if the fusion technique is not adapted to the information retrieval paradigm as well as to the TBIR and CBIR systems used. This means that the major challenge in information fusion is to find adjusted techniques for associating multiple sources of information for information retrieval. Since traditional work on multimodal integration has largely been heuristic-based, the understanding of how fusion works and by what it is influenced is limited This means that the major challenge in information fusion is to find adjusted techniques for associating multiple sources of information for either decision–making or information retrieval. From the other point of view, traditional work on multimodal integration has largely been heuristic-based. Still today, the understanding of how fusion works and by what it is influenced is limited. Therefore in next chapters, we will present a formal presentation for multimodality fusion in IR systems and then will propose a new method to combine different modalities in CBIR systems that can eliminate this limitation and will evaluate its effectiveness on medical images.

# CHAPTER FOUR
# INTEGRATED COMBINATION RETRIEVAL

As mentioned in pervious chapter, existence of some drawbacks and limits in multimodal fusion methods cause that some fusion methods even degrade the performance of total system. Therefore, we decided to show how overall system performance can be improved with combination of multimodality approach and how modalities should be combined. In this chapter, we first develop a formal model for multimodal fusion system based on set theory and then demonstrate the optimal status of such systems. Next, we proposed a new integrated method to combine different modalities in multimedia systems, called Integrated Combination retrieval and to investigate impact of this method, develop a CBIR system on medical images based on textual and visual modalities.

## 4.1 Formal Presentation of Multimodal Information System

A multimodal CBIR system can be considered as a scoring system, F, with a score function $f_s(d,q)$ that assigns each document d in the document set $D$ of $n$ documents, $D = \{d_1, d_2, .., d_n\}$, a real number which is the measure value of its relevance to a query $q$ in query set $Q$. Since different IR models generate quite different ranges of relevance scores, scores assigned by each model should be normalized before the combination. Thus, each score function $f_s(d,q): D \times Q \rightarrow R$ can be transformed to $\acute{f}_s(d,q): D \times Q \rightarrow [0\,,1]$ .

The set $Rel(q)$ is defined as documents that are identified by user or expert as a relevant document to query $q$.

The retrieved document of arbitrary modality i in response to query $q \epsilon Q$ , $Ret_i(q)$ is defined as follows:

$$Ret_i(q) = \{d_1, d_2\,, \dots,\; d_l\,\}$$

where

$$\acute{f}_s(d_k, q) \geq \; \acute{f}_s(d_{k+1}, q)\,,\; 1 \leq \; k \leq l - 1\;.$$

In the most of pervious literature of multi-modality fusion systems, the set of relevant retrieved document of fused m modality in response to query $q \epsilon Q$, $R_{mixed}(q)$ , was defined as follows*:*

$$R_{mixed}(q) = \left\{ d_k \epsilon\, D \mid d_k\, \epsilon \bigvee_{i=1}^{m} Ret_i(q) \right\}$$

It means that the best performance of multi-modality fusion appears when we can put all relevant retrieved documents of all fused modalities in response to a query, into relevant retrieved document set of combined modalities.

$$R^{*}{}_{mixed}(q) = \left\{ d_k \epsilon\, D \mid d_k\, \epsilon \bigcup_{i=1}^{m} Ret_i(q) \right\}$$

As illustrated in Fig.1, it seems that the best performance of multi-modality fusion appears when we can put all relevant retrieved documents of all fused modalities in response to a query *(*$Rel \cap (Ret_t \cup Ret_v)$*),* into relevant retrieved document set of fused modalities *(* $Rel \cap Ret_{mixed}$*).* More formally, Union of modalities' relevant retrieved document sets can achieve the best result in fusion. However, as obviously illustrated in figure 1, there are some documents that are relevant but not appear in any individual modalities' result set, which is the part that cannot be achieved with fusion techniques. This is the limit of multimodality fusion. Because of that, in literature of multimodal data fusion, some of the authors claimed that data fusion algorithms are competitive in performance and is not devoid of risks and sometimes can degrade the retrieval performance (Müller, H., Clough, P., Deselaers, T. and Caputo, B. (Eds.), 2010) (Wu, S., McClean, S. , I., 2006) . To overcome this issue, in our proposed approach theses missed documents could be placed in $R_{mixed}$ because their obtained similarity score,$\acute{f}_{mixed}(d, q)$, had been greater than similarity score of any documents in $R_t$ *or* $R_v$*,* due to linear combination method. Therefore It is obvious that our suggested approach could improve the number of relevant retrieved documents in response to an arbitrary $q \epsilon Q$ and effectiveness of whole system performance consequently. So, the optimal target of integrated multi-modality retrieval system realized when it could be defined as:

$$R^{*}{}_{mixed}(q) = \left\{ d_k \epsilon\ D \mid d_k\ \epsilon \bigcup_{i=1}^{m} Ret_i(q) \right\} \bigcup \left\{ d_k \epsilon\ D \mid d_k\ \epsilon \left( Rel(q) - \bigcup_{i=1}^{m} Ret_i(q) \right) \right\}$$

In Fig.1, *Ret<sub>t</sub>*, *Ret<sub>v</sub>*, *Ret<sub>mixed</sub>* are retrieved document sets for text, visual modalities and mixed of them, respectively. *Rel* is the relevant set for given query.

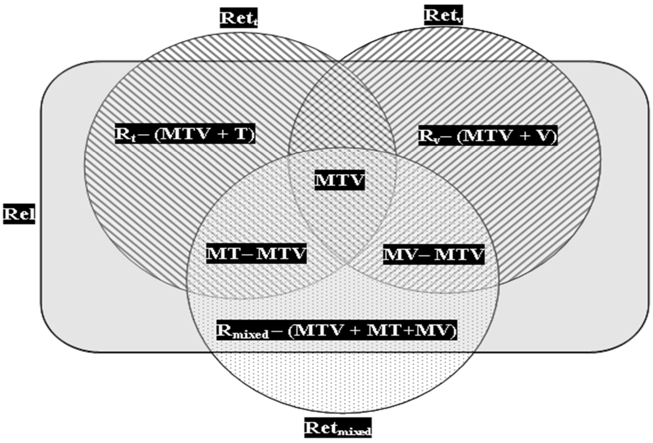

Figure 4 - 1. Diagram of relevant retrieved document set of different modalities in Integrated Combined Multimodal Retrieval

Our proposed method is a super level of late fusion because it can applied on both, similarity scores or ranks, of each modality feature that processed individually like as late fusion. The significant difference between this approach and late fusion is scale of combination. In our method, all of documents in data collection involve on combination. In contrast to late fusion that the number of combined document in list depends on value of a threshold based on score or rank. This is main reason that we call this method as “Integrated Combination”.

In order to evaluate the impact of our proposed method on improvement of overall system performance, we designed and implemented an experimental content based image retrieval system with two modalities, text and image. Textual modality was preprocessed and indexed in the Text Based Indexing and Retrieval subsystem In the second step, CBIR subsystem worked over the set of visual features of images in data collection and topic images. Finally, the third stair was multi-modality fusion subsystem and it combined the results obtained from different modalities in response to each query based on selected combination methods and produced the final result set of retrieved documents. In this phase, we implemented linear weighted method to combine different modalities result set based on both rank and score. Before score–based combinations, we normalized the similarity scores using Min-Max normalization (Lee, 1995) which is not the case of the rank–based strategies.

We performed our experiments with CLEF 2011 medical image classification and retrieval tasks dataset. The database includes 231,000 images from journals of BioMed Central at the PubMed Central database associated with their original articles in the journals. Beside, a single XML file is provided as textual metadata for all documents in the collection. Meanwhile, 30 topics, ten topics each for visual, textual and mixed retrieval, were chosen to allow for the evaluation of a large variety of techniques. Each topic has both a textual query and at least one sample query image. (Kalpathy-Cramer, 2011)

```
<TOPIC>

<ID>1</ID>

<TYPE>visual</TYPE>

<EN-DESCRIPTION>photographs of benign or malignant skin lesions.</EN-DESCRIPTION>

<DE-DESCRIPTION>Fotos von gutartigen oder bösartigen Melanomen.</DE-DESCRIPTION>

<FR-DESCRIPTION>des images de lésions de la peau bénignes ou malignes.</FR-DESCRIPTION>

</TOPIC>
```

Figure 4 - 2. The sample XML file of topics in data collection

```
<DOC>

<DOCNO>1471-213X-4-16-2</DOCNO>

<CAPTION> Identification of septal, outflow tract, and aortic arch malformations using multi-embryo MRI  ( a
–  e' ) Images of transverse sections from 5  Cited2 -/-  embryos obtained using the multi-embryo technique (
a – e ) compared with images from the same embryos obtained subsequently using the single embryo
technique ( a' – e' ). Section ( a, a' ) showing left and right atria and ventricles (la, ram, live, rave). The atria
are separated by the primary atria septum (pas), which is deficient at its ventral margin creating an osmium
premium type of atria septal defect (ASD-P). Section ( b, b' ) showing a ventricular septal defect (VSD) in the
interventricular septum (ivs). Section ( c, c' ) showing double outlet right ventricle, wherein the ascending
aorta (a-ao) and the pulmonary artery (pa) both arise from the right ventricle
</CAPTION>

<ARTICLEURL>http://www.biomedcentral.com/1471-213X/4/16</ARTICLEURL>

<ARTICLEFILENAME>10.1186_1471-213X-4-16.xml</ARTICLEFILENAME>

</DOC>
```

Figure 4 - 3.  The sample XML file for textual metadata of each image

Details of the phases processed modalities of system can describe as following:

## 4.2 Text Modality

In order to simplify the work, we split the XML file for textual metadata and represented each image in the collection as a structured document of xml file. We used Terrier IR Platform API, open source search engine written in Java and is developed at the School of Computing Science, University of Glasgow (Ounis, 2006), for our Text Based Information Retrieval subsystem. Terrier provides both efficient and effective search methods for large-scale document collections. To introduce flexibility to the processing and transformation of textual information, it requires a preprocessing in different ways. The order in which transformations were applied is as follows: 1) special characters deletion: characters with no meaning, like punctuation marks or blanks, are all eliminated; 2) stop words removal: discarding of semantically empty words, very high frequency words, 3) token normalization: converting all words to lower case 4) stemming: we used the Porter stemmer (Porter, 1980) as a process for removing the commoner morphological endings from words in English.

The indexing was done automatically by Terrier in a four stage process as follows: 1) handling of documents collection, 2) parsing each individual document, 3) processing of terms from documents, and 4) storing the index data structures. Terrier was designed to allow many different ways of indexing a corpus of documents, and this required some configuration about indexing fields and parameters.

The core functionality of retrieval phase was matching documents to queries and ranking documents. Matching employed a weighting model to assign a score to each of the query terms in a document. Since choice of the weighting model may crucially affect the performance of any information retrieval system, here we introduce some of famous weighting models that implemented in Terrier:

***TF-IDF weight***: The term frequency-inverse document frequency weight is a numerical statistic which reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. The term count in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term $\mathrm{t}$ within the particular document d. Thus we have $(\mathrm{t}, \mathrm{d})$ . The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$idf(t, D) = log \frac{|D|}{|\{d \in D : t \in d\}\,|}$$

where $|\mathrm{D}|$ is cardinality of D, or the total number of documents in the corpus and $|\{\mathrm{d} \in \mathrm{D} : \mathrm{t} \in \mathrm{d}\}\,|$ is the number of documents where the term $\mathrm{t}$ appears (i.e. $\mathrm{tf}(\mathrm{t}, \mathrm{d}) \neq 0$) If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the formula to $1 + |\{\mathrm{d} \in \mathrm{D} : \mathrm{t} \in \mathrm{d}\}\,|$. Mathematically the base of the log function does not matter and constitutes a constant multiplicative factor towards the overall result. Then the TF-IDF weight is calculated as

$$tf - idf(t, d, D) = tf(t, d) \times idf(t, D)$$

A high weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the idf's log function is always greater than 1, the value of idf (and tf-idf) is greater than 0. As a term appears in more documents then ratio inside the log approaches 1 and making idf and tf-idf approaching 0. If a 1 is added to the denominator, a term that appears in all documents will have negative idf, and a term that occurs in all but one document will have an idf equal to zero.

***Okapi BM25***: The name of the actual ranking function is BM25that used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Robertson and others. BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity). One of the most prominent instantiations of the function is as follows:

Given a query Q, containing keywords $\mathrm{q_1, q_2, \ldots, q_n}$ , the BM25 score of a document D is:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i)\ .\ \frac{f(q_i,D)\,.\,(k_1+1)}{f(q_i,D) + k_1\,.\left(1 - b + \left(b\,.\frac{|D|}{avg\ dl}\right)\right)}$$

where $\mathrm{f(q_i, D)}$is term frequency of $\mathrm{q_i}$ in the document D, $\mathrm{|D|}$ is the length of the document D in words, and $\mathrm{avg\,dl}$ is the average document length in the text collection from which documents are drawn. $\mathrm{k_1}$and b are free parameters, usually chosen, in absence of an advanced optimization, as $\mathrm{k_1} \in [1.2\,, 2.0]$ and $\mathrm{b = 0.75}$ [1]. $\mathrm{IDF(q_i)}$ is the inverse document frequency weight of the query term $\mathrm{q_i}$ . It is usually computed as:

$$IDF(q_i) = \ log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

where N is the total number of documents in the collection, and $\mathrm{n(q_i)}$ is the number of documents containing $\mathrm{q_i}$.

We compared performance of the subsystem using variety of implemented weighting models in Terrier. Although BM25 and TF- IDF weighting model had been evaluated as the most effective weight model according to our evaluation, but we chose DFR-BM25 weighting model (Amati, 2003) as base textual modality of our system because its result was almost the average values between results of other weighting model. After that we calculate the similarity score of all documents in collection corresponding to each query topic and then sort them in descending order as ranked list.

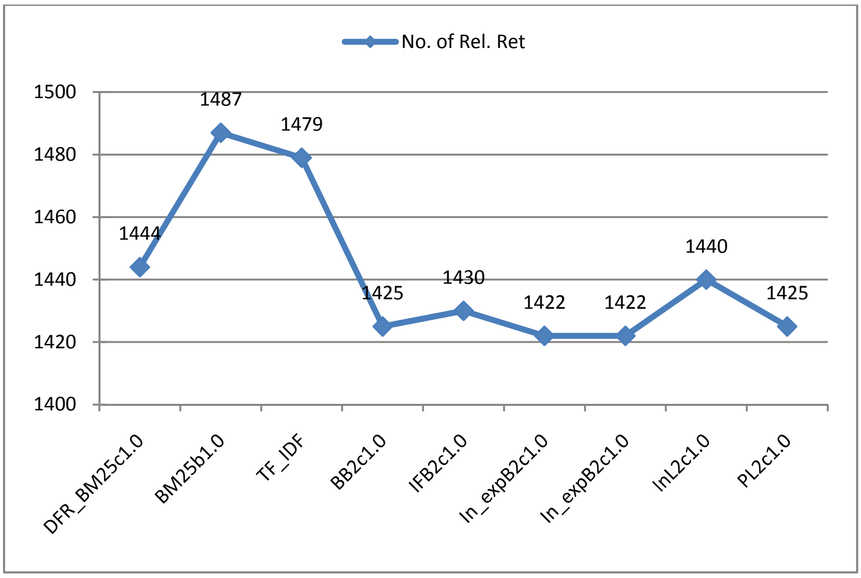

Figure4 - 4. Comparison on performance of different textual weighting model based on number of relevant retrieved documents

## 4.3 Visual modality

We extracted features for all images in test collection and query examples using Rummager tool (Chatzichristofis S. A., 2009), which is developed in the Automatic Control Systems & Robotics Laboratory at the Democritus University of Thrace-Greece. Since selection of right features is the major aspect to attain discriminative and sufficient retrieval systems, then we examined the performance of all extracted feature and perceived that compact composite features like CEDD and FCTH have satisfactorily retrieval result on our image collection and require noticeably lower computational power and storage space.

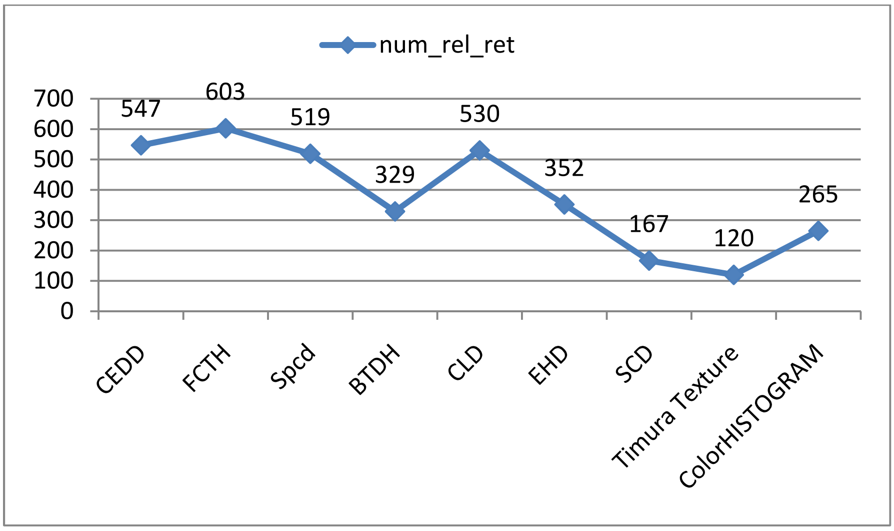

Figure4 - 5. Comparison on performance of different low level features

Table 4 - 1. Comparison on performance of different low level features

| Feature name | CEDD | FCTH | SpCD | BTDH | CLD | EHD | SCD |
|---|---|---|---|---|---|---|---|
| No. of Rel.Ret | 547 | 603 | 519 | 329 | 530 | 352 | 167 |
| MAP | 0.018 | 0.0186 | 0.0127 | 0.0059 | 0.0126 | 0.0116 | 0.0013 |
| Rprec | 0.046 | 0.0405 | 0.0305 | 0.0181 | 0.0352 | 0.0268 | 0.0052 |
| Bpref | 0.0633 | 0.0755 | 0.0707 | 0.0496 | 0.0651 | 0.0556 | 0.0318 |

Therefore we chose CEDD as base visual modality of our system. Also, we assessed performance of different similarity function on Compact Composite features and comprehended Euclidean distance on CEDD and FCTH features and Cosine distance function on SPCD and BTDH produce the best performance. Then we calculate the similarity between query and dataset objects using Euclidean distance in matching phase. Then we sorted all of dataset images in a descending list based on the value of similarity score in corresponding to each query example image.

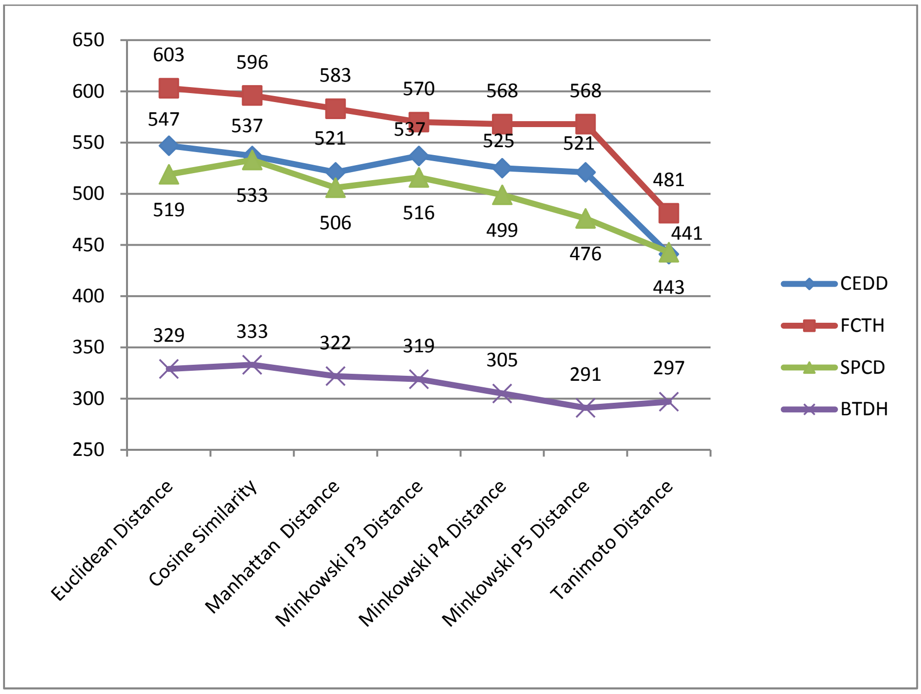

Figure4 - 6. Distance function performance evaluation of different features based on number of relevant retrieved documents.

# CHAPTER FIVE
# EXPERIMENTATIONS

To appraise the performance of different combination approaches on the result set of different features, we set up a set of experiment.

- *Early Fusion*: We used feature concatenation method on the synchronous compact composite feature vectors of all images in data collection and topics and concatenated them as the joint feature vector. We used Euclidean distance for similarity measure and selected top 1000 documents for each query.
- *Late Fusion with Substitution Value of Zero (LFSVZ)*: We applied CombSUM function on similarity scores of first 1000 top retrieved documents of each feature result set, response to each query using Fagin’s A0 combination algorithm (Fagin, R., 1999). In this phase, we normalized the similarity scores using Min-Max normalization function (Lee, J. H., 1995) before combination. According to Fagin’s A0 combination algorithm, we substitute zero as similarity score of documents that are not appeared in retrieved document list.

Table5-1 presents performance comparison of above mentioned methods on combination of compact composite features. This experiment performed based on number of relevant retrieved documents and mean average precision (MAP). As it clearly obvious, ICMR outperforms any of the other fusion methods in terms of both measure.

Table 5 - 1. Comparison of different combination method performance.

| Combination Function | Combination method | # of Relevant Retrieved | MAP |
|---|---|---|---|
| CombSUM(CEDD,FCTH) | Early Fusion | 658 | 0.0201 |
| | LFSVZ | 643 | 0.0194 |
| | ICMR | 665 | 0.199 |
| CombSUM (CEDD,FCTH,SpCD) | Early Fusion | 676 | 0.0231 |
| | LFSVZ | 541 | 0.0198 |
| | ICMR | 699 | 0.0252 |

In order to assess the performance of our suggested method on multimodal CBIR system, we set some experiments on our base textual and visual modality. In these experiments, we applied weighted CombSUM function using different values to weight for textual modality and value of 1 for visual modality's weight. As illustrated in Table 5-2, it is obvious that our integrated combination method performs better than another method in all corresponding weighting schemes based on similarity score. There are some documents in result set of ICMR that do not appear in none of combined modalities' result sets because their obtained similarity scores had been greater than similarity score of any documents in individual modalities after combination. For more detail, let's consider Table 5-3 that illustrates some relevant retrieved documents in response to query #18. The first data row demonstrates threshold of normalized similarity score in top 1000 retrieved documents. It is apparent that similarity scores of these documents was less than thresholds in both modalities and they did not appear in top 1000 retrieved documents list of modalities. But due to ICMR using weighted CombSUM, their obtained scores placed in top 1000.

We also found that in medical image data collections, when weight of textual modality is about 1.7 folds of visual modality, performance of ICMR is more effective in score based approach.

Table 5 -2. Performance comparison of different combination methods on similarity scores of modalities using different weights for textual modality

| | Text Weight | # Rel-Ret | Map | Rprec | Bpref | p@5 | p@100 |
|---|---|---|---|---|---|---|---|
| ICMR | 0.7 | 1454 | 0.2015 | 0.2574 | 0.2515 | 0.4933 | 0.2037 |
| | 1 | 1578 | 0.2289 | 0.2784 | 0.2744 | 0.4933 | 0.218 |
| | 1.5 | 1597 | 0.2372 | 0.2881 | 0.2738 | 0.4733 | 0.221 |
| | 1.7 | 1599 | 0.2341 | 0.2873 | 0.2704 | 0.4667 | 0.2217 |
| | 2 | 1595 | 0.2307 | 0.2706 | 0.2606 | 0.4533 | 0.2177 |
| | 2.5 | 1573 | 0.2293 | 0.2537 | 0.2501 | 0.44 | 0.2167 |
| | 3 | 1559 | 0.2269 | 0.2561 | 0.2462 | 0.42 | 0.215 |
| LFSVZ | 0.7 | 560 | 0.0754 | 0.1175 | 0.1198 | 0.3333 | 0.0997 |
| | 1 | 720 | 0.0865 | 0.1315 | 0.137 | 0.36 | 0.117 |
| | 1.5 | 1246 | 0.1498 | 0.2059 | 0.1981 | 0.3933 | 0.1843 |
| | 1.7 | 1385 | 0.1659 | 0.2124 | 0.2071 | 0.4 | 0.19 |
| | 2 | 1484 | 0.178 | 0.2188 | 0.2182 | 0.4067 | 0.1913 |
| | 2.5 | 1515 | 0.1918 | 0.2293 | 0.2249 | 0.4067 | 0.193 |
| | 3 | 1479 | 0.1964 | 0.2401 | 0.2314 | 0.4067 | 0.196 |

Table 5 - 3. Details of ICMR in response to query #18

| | Text | Visual | Mixed |
|---|---|---|---|
| Threshold in 1000th top score | 0.4053 | 0.8677 | 0.6486 |

| Document ID | Similarity Scores | | |
|---|---|---|---|
| | Textal | Visual | Mixed |
| 1471-213X-4-16-2 | 0.3029 | 0.7768 | 1.2919 |
| 1471-213X-4-16-3 | 0.3242 | 0.8055 | 1.3567 |
| 1471-213X-4-16-5 | 0.3223 | 0.7837 | 1.3318 |

But our finding in rank based approach was completely in difference. In this approach, performance of ICMR was worth than LFSVZ. Details are mentioned in Table 5-4. In ranked based ICMR, improvement on system performance correlated with increasing of textual modality’s weight. While in LFSVZ, growth in textual modality’s weight decrease the effectiveness of system in case of MAP and number of relevant retrieved documents.

Table 5 - 4. Performance comparison of different combination methods on rank of modalities using different weights for textual modality

| | Text Weight | # Rel-Ret | Map | Rprec | Bpref | p@5 | p@100 |
|---|---|---|---|---|---|---|---|
| ICMR | 0.7 | 999 | 0.0113 | 0.0119 | 0.104 | 0 | 0.0097 |
| | 1 | 1025 | 0.0115 | 0.0057 | 0.1048 | 0.02 | 0.0043 |
| | 1.5 | 1067 | 0.0119 | 0.0041 | 0.1151 | 0.0067 | 0.0057 |
| | 1.7 | 1084 | 0.0123 | 0.0061 | 0.1153 | 0.0133 | 0.008 |
| | 2 | 1100 | 0.0128 | 0.0074 | 0.1162 | 0.0067 | 0.007 |
| | 2.5 | 1118 | 0.0128 | 0.01 | 0.1145 | 0 | 0.0067 |
| | 3 | 1135 | 0.0134 | 0.0073 | 0.1184 | 0 | 0.0093 |
| LFSVZ | 0.7 | 1413 | 0.0202 | 0.0157 | 0.1064 | 0.0067 | 0.0197 |
| | 1 | 1394 | 0.0216 | 0.0216 | 0.1119 | 0.0267 | 0.0203 |
| | 1.5 | 1345 | 0.0196 | 0.0176 | 0.1139 | 0 | 0.0203 |
| | 1.7 | 1329 | 0.0196 | 0.0183 | 0.1108 | 0.0333 | 0.0187 |
| | 2 | 1308 | 0.019 | 0.0244 | 0.1129 | 0.0133 | 0.0207 |
| | 2.5 | 1270 | 0.0195 | 0.0229 | 0.1222 | 0.0333 | 0.0237 |
| | 3 | 1227 | 0.0176 | 0.0224 | 0.1135 | 0.0267 | 0.022 |

To study the impact of our method on improvement of performance of multi-modality information retrieval in depth, we analyzed our experiments based on formal presentation of integrated combination mentioned in pervious chapter. Result of this examination confirmed our claim about impact of ICMR on improvement of combination system performance in score based approach.

Table 5 - 5. Impact of combination function on different modalities in ICMR

| Text Weight | $\boldsymbol{R_{mixed}}$ | $\boldsymbol{MV} = R_{mixed} \cap R_v$ | $\boldsymbol{MT} = R_{mixed} \cap R_t$ | $\boldsymbol{VT} = R_v \cap R_t$ | $\boldsymbol{MTV} = R_{mixed} \cap R_t \cap R_v$ | $\boldsymbol{MT - MTV}$ | $\boldsymbol{MV - MTV}$ | $\boldsymbol{R_t - (MTV + T)}$ | $\boldsymbol{R_v - (MTV + V)}$ | $\boldsymbol{R_{mixed} - (MTV + MT + MV)}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1559 | 287 | 1410 | 219 | 219 | 1191 | 68 | 34 | 260 | 81 |
| 2.5 | 1573 | 298 | 1404 | 219 | 219 | 1185 | 79 | 40 | 249 | 90 |
| 2.0 | 1595 | 310 | 1393 | 219 | 219 | 1174 | 91 | 51 | 237 | 111 |
| **1.7** | **1599** | 321 | 1373 | 219 | 219 | 1154 | 102 | 71 | 226 | 124 |
| 1.5 | 1597 | 329 | 1354 | 219 | 219 | 1135 | 110 | 90 | 218 | **133** |
| 1.0 | 1578 | 394 | 1254 | 219 | 219 | 1035 | 175 | 190 | 153 | 149 |
| 0.7 | 1454 | 432 | 1089 | 219 | 219 | 870 | 213 | 355 | 115 | 152 |

Zero values in the last column of table 5-6 shows that no document out of modalities' result set are not included in final relevant retrieved document set in LFSVZ while there are some document in ICMR final result set that do not appear in any combined modalities results. Moreover regardless to weights of linear function, in all experiments, retrieved documents that evaluate as relevant in both of textual and visual modalities are retained in relevant retrieved document of combined set.

Table 5 - 6. Impact of combination function on different modalities in LFSVZ

| Text Weight | R<sub>mixed</sub> | $\boldsymbol{MV} = R_{mixed} \cap R_v$ | $\boldsymbol{MT} = R_{mixed} \cap R_t$ | $\boldsymbol{VT} = R_v \cap R_t$ | $\boldsymbol{MTV} = R_{mixed} \cap R_t \cap R_v$ | $\boldsymbol{MT - MTV}$ | $\boldsymbol{MV - MTV}$ | $\boldsymbol{R_t - (MTV + T)}$ | $\boldsymbol{R_v - (MTV + V)}$ | $\boldsymbol{R_{mixed}} - (\boldsymbol{MTV + MT + MV})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1479 | 261 | 1437 | 219 | 219 | 1218 | 42 | 7 | 286 | 0 |
| 2.5 | 1515 | 337 | 1397 | 219 | 219 | 1178 | 118 | 47 | 210 | 0 |
| 2.0 | 1484 | 429 | 1274 | 219 | 219 | 1055 | 210 | 170 | 118 | 0 |
| 1.7 | 1385 | 481 | 1123 | 219 | 219 | 904 | 262 | 321 | 66 | 0 |
| 1.5 | 1246 | 506 | 959 | 219 | 219 | 740 | 287 | 485 | 41 | 0 |
| 1.0 | 720 | 543 | 396 | 219 | 219 | 177 | 324 | 1048 | 4 | 0 |
| 0.7 | 560 | 547 | 232 | 219 | 219 | 13 | 328 | 1212 | 0 | 0 |

# CHAPTER SIX
# CONCLUSION

In this thesis, we investigate to find appropriate combination of textual and visual modalities in Content-based Medical Image Retrieval systems. In CBIR systems, merging result-sets of different modalities is crucial, since an effective combination of the different modalities directly influence the overall performance of retrieval systems, the experiments of this study originated from our participation at ImageCLEF 2011 Medical Image Retrieval Track where we had received the best five rank in mixed retrieval run of textual and visual modalities.

In order to investigate on possible combinations methods, we first do some evaluations to determine the appropriate low-level features and distance functions on visual modality. Then we presented an in depth investigation on different combination methods for multimodal CBIR systems. In this way, we show how overall system performance can be improved with combination of multimodality approach and how modalities should be combined. We also suggest a new combination approach which is based on integrating multimodal retrieval. We also compared this model with common combination methods for multimodal content-based medical image retrieval in pervious literature.

Several major findings of this study we gained from experimentations can be summarized as follows:

- We show that effective combination of textual and visual modalities improves the overall performance of Content-based Medical Image Retrieval Systems.
- It is clear that integrated retrieval outperforms all fusion techniques in score based approach, regardless of late or early fusion, in multimodal CBIR systems.
- In the best combination of textual and visual modalities, weight for textual modality is about 1.7 folds of visual modality weight.
- Common documents in relevant retrieved set of different modalities also appears in relevant retrieved document set of combined modality too, regardless weights or methods.

- Limitation of fusion methods originates from their restrictions of combination situation such as number of document participated in combination method.
- We show that extracting compact composite image feature as visual modality can improve the effectiveness of medical content-based image retrieval systems.
- We found that Euclidean distance function on CEDD and FCTH features and Cosine distance function on SPCD and BTDH gives the best performance.

Our study can be extended in several ways, in future. First, it would be good to apply this experimentation results to other medical image collection and verify that our findings produces the similar results on similar CBIR systems working on textual and visual modalities. Second, the impact of other normalization methods and similarity functions on system performance can also be further investigated. Lastly, our study can be extended into other domain of CBIR Systems rather than medical domain.

## REFERENCES

Amati, G. (2003). Probabilistic Models for Information Retrieval based on Divergence from Randomness. School of Computing Science, University of Glasgow : PhD Thesis.

Atrey P. K., Hossain, M. A., Saddik, A. E., Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey . Springer Multimedia Systems Journal, *16*(6), 345-379.

Blaser, A. (1979). Database Techniques for Pictorial Applications. Lecture Notes in Computer Science.

Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., Laere, J. V., Niklasson, L., Nilsson, M., Persson, A., Ziemke, T. (2007). On the Definition of Information Fusion as a Field of Research. Technical, University of Skovde, School of Humanities and Informatics.

Burkhardt, H. ,. (2000). "Invariant features for discriminating between equivalence classes Nonlinear Model-based Image Video Processing and Analysis. John Wiley and Sons.

Cawkill, A. E. (1993). The British Library's Picture Research Projects: Image, Word, and Retrieval. Advanced Imaging, *8*(10), 38-40.

Chang, N. S. (1979). A relational database system for images. Purdue University.

Chang, N. S. (1980). Query by pictorial example . IEEE Trans. on Software Engineering, 6(6), 519-524.

Chang, S. K. (1981). Pictorial database systems. IEEE Computer Magazine, *14*(11), 13-21.

Chang, T. K. (1993). Texture analysis and classification with tree-structured wavelet transform. IEEE Trans. on Image Processing, *2*(4), 429-441.

Chatzichristofis, S. A. (2008). CEDD: Color and Edge Directivity Descriptor – a compact descriptor for image indexing and retrieval. 6th International Conference in advanced research on Computer Vision Systems (ICVS),Santorini, Greece.

Chatzichristofis, S. A. (2008). FCTH: Fuzzy Color and Texture Histogram- a low level feature for accurate image retrieval. 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), IEEE Computer Society, Klagenfurt, Austria.

Chatzichristofis, S. A. (2009). IMG(RUMMAGER): AN INTERACTIVE CONTENT BASED IMAGE RETRIEVAL SYSTEM. 2nd International Conference on Similarity Search and Applications (SISAP) (pp. 151-153). Prague, Czech Republic.: IEEE Computer Society.

Chatzichristofis, S. A. (2010). ACCURATE IMAGE RETRIEVAL BASED ON COMPACT COMPOSITE DESCRIPTORS AND RELEVANCE FEEDBACK INFORMATION. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), *24*(2), 207-244.

Chatzichristofis, S. A. (2010). CONTENT BASED RADIOLOGY IMAGE RETRIEVAL USING A FUZZY RULE BASED SCALABLE COMPOSITE DESCRIPTOR. Multimedia Tools and Applications, *46*(2-3), 493-519.

Chieu, H. L. (2004). Query based event extraction along a timeline. International ACM Conference on Research and Development in Information Retrieval, (pp. 425–432). Sheffield.

Datta, R. J. (2008, April). Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, *40*(2).

Datta, R., Joshi, D., Li, J., Wang, J. Z. (2008, April). Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, *40*(2).

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. on Information Theory, *36*, 961-1005.

Dimitrovski, I., Gorgevik, D., Loskovska, S. (2007). Web-based medical image retrieval system. Proceedings ofInformation Society – IS2007, (pp. 19-22). Ljubljana SLOVENIA.

Dowe, J. (1993). Content-based retrieval in multimedia imaging. SPIE Storage and Retrieval forImage and Video Database.

Fagin, R. (1999). Combining fuzzy information from multiple systems. Journel of Computer and Systems Sciences, 83-99.

Foley, J. D. (1990). Computer graphics: principles and practice. Addison-Wesley.

Furht, B. S. (1995). Video and Image Processing in Multimedia Systems. Kluwer Academic Publishers.

Hu, R. R. (2008). Dissimilarity measures for content-based image retrieval. IEEE Int. Conf. on Multimedia and Expo, (pp. 1365-1368).

Huang, J. ,. (1997). Image indexing using color correlogram. IEEE Int. Conf. on Computer Vision and Pattern Recognition, (pp. 762-768).

ISO/IEC/JTC1/SC29/WG11. (2000). Core Experiment Results for Edge Histogram Descriptor. MPEG Document , Beijing.

Jain, A. K. (1989). Fundamental of Digital Image Processing. Englewood Cliffs, Prentice Hall.

Jain, A. K. (1991). Unsupervised texture segmentation using Gabor filters. Pattern Recognition, *24*(12), 1167-1186.

Jain, R. (1992). Proc. US NSF Workshop Visual Information Management Systems.

Kalpathy-Cramer, J. M. (2011). Overview of the CLEF 2011 Medical Image Classification and Retrieval Tasks. CLEF (Notebook Papers/Labs/Workshop).

Kashyap, R. L. (1983). Estimation and Choice of neighbors in Spatial-Interaction Models of Images. " , IEEE transactions on information theory, IT-29(1), 60-72.

Kilic, D., Alpkocak, A. (2011). An Expansion and Reranking Approach for Annotation-based Image Retrieval from Web. Expert Systems With Applications, *38*(10), 13121-13127.

Lee, J. H. (1995). Combining multiple evidence from different properties of weighting schemes. 18th annual international ACM SIGIR conference on research and development in information retrieva (pp. 180–188). ACM press.

Lee, J. H. (1995). Combining multiple evidence from different properties of weighting schemes. 18th annual international ACM SIGIR conference on research and development in information retrieva (pp. 180–188). ACM press.

Liu, F. P. (1996). Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE Trans. on Pattern Analysis and Machine Learning, 18*(7).

Llinas J., Bowman, Ch., Rogova, G., Steinberg, A., Waltz, E., White, F. (2004). Revisiting the JDL Data Fusion Model II. 7th International Conference on Information Fusion, (pp. 1218-1230).

Ma, W. Y. (1995). A comparison of wavelet features for texture annotation. Proc. Of IEEE Int. Conf. on Image Processing, 2, pp. 256-259. Washington D.C.

Manjunath, B. S. (1996). Texture features for browsing and retrieval of image data. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 18*(8), 837-842.

Mao, J. (1992). Texture classification and segmentation using multi resolution simultaneous autoregressive models. Pattern Recognition, *25*(2), 173-188.

Mathias, E. (1998). Comparing the influence of color spaces and metrics in content-based image retrieval. Proceedings of International SSymposium on Computer Graphics, Image Processing, and Vision, (pp. 371-378).

Muller, H. C. (2010). ImageCLEF. Springer Verlog .

Müller, H., Clough, P., Deselaers, T. and Caputo, B. (Eds.). (2010). ImageCLEF, The Information Retrieval Series 32. Berlin Heidelberg: Springer-Verlag.

Niblack, W. e. (1993). Querying images by content, using color, texture, and shape. SPIE Conference on Storage and Retrieval for Image and Video Database, 1908, 173-187.

Ounis, I. A. (2006). Terrier: A High Performance and Scalable Information Retrieval Platform. ACM SIGIR'06 Workshop on Open Source Information Retrieval.

Pass, G. (1996). Histogram refinement for content-based image retrieval. IEEE Workshop on Applications of Computer Vision, (pp. 96-102).

Poh, N. B. (2005). How do correlation and variance of baseexperts affect fusion in biometric authentication tasks. *IEEE Trans. Signal Process, 53*, 4384-4396.

Porter, M. F. (1980). An algorithm for suffix stripping. *Electronic library and information systems, 14*(3), 130-137.

Rahman, M. M. (2004). Medical Image Retrieval and Registration: Towards Computer Assisted Diagnostic Approach. Proceedings of the IDEAS Workshop on Medical Information Systems: The Digital Hospital Issue, (pp. 78-89).

Rui, Y. H. (1999). Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation, 10,* 39-62.

Sanderson, C. P. (2004). Identity verification using speech and face information. *Digital Signal Processing*, *14*(5), 449-480.

Smeulders, A. M. (2000). Content-based image retrieval at the end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence, *22*(12), 1349-1380.

Snoek C., Worring, M., Smeulders, A. (2005). Early versus late fusion in semantic video analysis. ACM Multimedia, (pp. 399-402). Singapore.

Snoek, C. W. (2005). Early versus late fusion in semantic video analysis. ACM International Conference on Multimedia, (pp. 399–402). Singapore.

Stricker, M. O. (1995). Similarity of color images. SPIE Storage and Retrieval for Image and Video Databases III, 2185, 381-392.

Swain, M. J. (1991). Color indexing. International Journal of Computer Vision, *7*(1), 11-32.

Tamura, H. ,. (1984). Image database systems: A survey. Pattern Recognition, *17*(1), 29-43.

Tamura, H. M. ( 1978.). "Texture features corresponding to visual perception. IEEE Trans. On Systems, Man, and Cybernetics, Smc-8(6).

Wu, S. (2009). Applying statistical principles to data fusion in information retrieval. Expert Systems With Applications, 2997-3006.

Wu, S., McClean, S. , I. (2006). Performance prediction of data fusion for information retrieval. Inf. Process. Manage, *42*(4), 899-915.

Wu, Y. C. (2004). Optimal multimodal fusion for multimedia data analysis. ACM International Conference on Multimedia, (pp. 572–579). New York.

Wu, Z. C. (2006). Multi-level fusion of audio and visual features for speaker identification. International Conference on Advances in Biometrics, (pp. 493-499).