LES MIXTURES DE DIRICHLET ET LEURS APPORTS POUR LA CLASSIFICATION ET LA RECHERCHE D'IMAGES PAR LE CONTENU

par

Nizar Bouguila

mémoire présenté au Département de mathématiques et d'informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, novembre 2002

Le 15 Novembre 2002,
Date

le jury a accepté le mémoire de M. Nizar Bouguila dans sa version finale.

Composition du jury

Membre :	M. Djemel Ziou (Direction) Département de mathématiques et d'informatique
Membre :	M. Jean Vaillancourt (Codirection) Université du Québec à Hull
Membre :	M. Antoine Tabbonne INRIA
Membre et président-rapporteur :	M. Alain Boulanger Département de mathématiques et d'informatique

A ma chère mère et mon cher père.

A mes frères.

A mes proches.

SOMMAIRE

Le développement de la médecine moderne dans le domaine des techniques de diagnostic comme la radiologie, l'histopathologie et la tomographie avait comme résultat l'explosion du nombre et de l'importance des images médicales sauvegardées par la majorité des hôpitaux. Afin d'aider les médecins à confirmer leurs diagnostics, plusieurs systèmes de recherche d'images médicales ont vu le jour. La conception de ces systèmes présente plusieurs étapes. Nous pensons que le résumé des bases de données d'images est une étape importante dans chaque système de recherche. En effet, la catégorisation d'une base de données d'images facilite énormément la recherche et permet de localiser les images voulues en un minimum de temps.

Dans ce mémoire, nous étudions en un premier temps, les différents problèmes communs à tous les systèmes de recherche d'images à savoir l'indexation, l'extraction des caractéristiques, la définition des mesures de similarités et le retour de pertinence. Nous étudions aussi d'autres catégories de problèmes spécifiques à la recherche d'images. Cette étude est complétée par une analyse des systèmes existants les plus connus.

Dans la deuxième partie du mémoire, nous nous intéressons aux mixtures de Dirichlet et comment on peut les exploiter pour la classification, en particulier le résumé des bases de données d'images. Contrairement aux approches classiques qui considèrent la loi normale comme densité, nous utilisons une généralisation de la Dirichlet pour l'adapter plus aux problèmes réels. Notre approche est traduite par un modèle mathématique basé sur le maximum de vraisemblance et la méthode de Fisher. Une interprétation très intéressante de notre méthode, basée sur la statistique géométrique, est donnée. Finalement, nous présentons des évaluations contextuelles et non-contextuelles, qui prouvent la validité de notre méthode.

REMERCIEMENTS

J'aimerais d'abord remercier mon directeur de recherche, le professeur Djemel Ziou, pour m'avoir donné l'opportunité de faire cette maîtrise. J'aimerais le remercier pour toutes les discussions que nous avons eues et qui m'ont permis d'avancer dans mon travail, de même que pour tous les conseils qu'il m'a donnés. Je veux aussi remercier mon codirecteur de recherche, le professeur Jean Vaillancourt, pour ces idées indispensables et sa sympathie. J'aimerais remercier mes collègues du groupe MOIVRE (MOdélisation en Imagerie, Vision et REseaux de neurones) pour leur aide. Un gros merci à la Mission Universitaire de Tunisie en Amérique du nord, à Bell Canada ainsi qu'au département de mathématiques et d'informatique et à mes directeurs de recherche pour leur apport financier sans lequel je n'aurais pu faire cette maîtrise. J'aimerais finalement remercier mes parents, mes frères et Claudine Couture pour leur support moral tout au long de cette maîtrise.

Table des matières

Introduction

Durant les dernières années, l'intérêt accordé aux images digitales a augmenté énormément. Cet intérêt est stimulé, au moins, par l'évolution rapide de l'imagerie dans plusieurs domaines tels que le World Wide Web et la médecine. Les utilisateurs, dans plusieurs domaines professionnels, essaient d'exploiter au maximum les opportunités offertes pour manipuler ces images après y avoir accédé. Cependant, ils découvrent que la localisation d'une image désirée dans une base de données est une source considérable de frustration. Les méthodes classiques de recherche d'images, basées sur le texte, avaient deux difficultés majeures. La première est le travail considérable que demande l'annotation manuelle des images. La deuxième, qui est selon nous plus essentielle, est la richesse de l'image du point de vue contenu et la subjectivité de la perception humaine. En effet, une même image peut être vue de différentes façons par des personnes différentes. Cette subjectivité peut causer plusieurs imprécisions dans l'annotation et ensuite dans la recherche. Ces problèmes ont eu comme conséquence l'augmentation de l'intérêt accordé aux techniques de recherche d'images par le contenu.

Dans notre travail, nous nous intéressons aux problèmes relatifs à la recherche d'images médicales. Comme les images médicales ont des particularités, nous commençons par une étude sur ces images. Ensuite, nous étudions les différents points essentiels pour la conception d'un système de recherche d'images. D'abord, il y a l'identification et l'extraction des caractéristiques pertinentes des images, ainsi que la production des métadonnées qui peuvent aider à localiser facilement les images voulues. Ensuite, il y a l'interaction de l'utilisateur avec le système pour l'attribution d'un degré d'importance à chaque caractéristique ou bien pour le raffinement de la recherche. Il y a aussi le problème qui consiste à définir une mesure de similarité qui correspond mieux à la perception de

l'usager. D'autres problèmes en recherche d'images sont aussi importants telles que l'indexation et la classification qui permettent aux usagers de naviguer dans la collection d'images. Dans la majorité des systèmes que nous avons étudiés, l'indexation se fait en utilisant des structures de données de grande complexité à savoir les arbres. Cependant, cette approche devient très complexe pour une grande base de donnée (des milliers d'images). A notre avis, d'autres moyens statistiques sont plus efficaces non seulement du point de vue complexité mais aussi en ce qui concerne le temps de classification. Les mixtures sont l'un de ces moyens. Grâce aux mixtures, les données d'une base de données peuvent être partitionnées en catégories homogènes. Toutefois, la majorité des travaux existants utilise la loi normale dans les mixtures. Bien que cette loi offre plusieurs avantages, telles que sa nature isotropique et sa capacité à représenter une distribution juste par une moyenne et une matrice de covariance, elle échoue à donner de bons résultats lorsque la partition n'est pas gaussienne. Pour cette raison, nous avons considéré la distribution de Dirichlet qui offre une très grande flexibilité et qui permet plusieurs formes. Cette densité a été généralisée afin de s'adapter plus aux problèmes réels. Les paramètres de notre nouvelle mixture, basée sur la Dirichlet, sont estimés grâce au maximum de vraissemblance et la méthode de Fisher. De même, plusieurs évaluations contextuelles et non-contextuelles ont été faites pour valider notre méthode.

Ce mémoire est divisé en deux chapitres. Le premier chapitre comprend une étude analytique de la problématique de la recherche d'images particulièrement en médecine. L'article **An Overview of Medical Image Retrieval Systems** a fait l'objet d'un rapport d'activité. Dans le second chapitre, nous proposons une nouvelle mixture. L'article **Maximum Likelihood Estimation of the Generalized Dirichlet Mixture** est soumis.

Chapitre 1

La recherche d'images médicales : une synthèse

Dans la première partie de ce mémoire, nous présentons le travail intitulé **An Overview of Medical Image Retrieval Systems**. Ce travail concerne l'étude des problèmes relatifs aux systèmes de recherche d'images, particulièrement dans le domaine médicale. Plusieurs travaux de synthèse, concernant la recherche d'images en général, ont été faits [10, 9, 8, 7, 4]. Cependant, nous pensons qu'il vaut mieux traiter ce sujet pour des domaines spécifiques. En effet, les systèmes de recherche d'images dépendent énormément de la culture de leurs utilisateurs ainsi que des caractéristiques des données. La recherche d'images dans des domaines spécifiques, tels que le Web et l'image médicale, n'a pas fait l'objet de grands travaux de synthèse. Dans la littérature, nous avons trouvé un seul rapport de synthèse traitant de la recherche d'images sur le web [6]. Par contre, aucun rapport traitant de la recherche d'images médicales n'a été trouvé . Cela nous a motivés à effectuer une étude complète sur les systèmes de recherche d'images médicales.
Comme les images médicales ont des caractéristiques différentes des autres images, nous avons commencé par une étude sur les différents types d'images médicales utilisées. Ensuite, nous nous sommes attaqués aux problèmes relatifs à la recherche d'images médicales. Une grande partie de ces problèmes concerne la recherche d'images en général telle que la considération des spécificités des utilisateurs et de leurs besoins, l'extraction

des caractéristiques des images et la production des métadonnées, la définition des mesures de similarités, l'indexation, le retour de pertinence et les méthodes d'évaluation. Cette étude a été projetée sur quelques systèmes existants pour la recherche d'images médicales. Nous avons analysé ces systèmes en respectant différents points à savoir leurs objectifs, la formulation des requêtes, l'indexation et les caractéristiques utilisées.
Le sujet a été proposé par les Professeurs Djemel Ziou et Jean Vaillancourt. La recherche bibliographique ainsi que la synthèse des problèmes et l'analyse des systèmes existants ont été faites sous leur supervision. Cela a fait l'objet d'un rapport [2] qui apparaît dans les pages suivantes de ce mémoire.

An Overview of Medical Image Retrieval Systems [1]

N. Bouguila(1), D. Ziou(1) and J. Vaillancourt(2)

(1) DMI, Faculté des sciences
Université de Sherbrooke
Sherbrooke, Qc, Canada J1K 2R1.

(2) Université du Québec à Hull
283, boulevard Alexandre-Taché
Hull, Qc, Canada J8X 3X7.

Abstract

The increasing reliance of modern medicine on diagnostic techniques such as radiology, histopathology and computerized tomography has resulted in an explosion in the number and importance of medical images now stored by most hospitals. While the prime requirement for medical imaging systems is to be able to display images relating to a patient, there is an increasing interest in the use of CBIR (Content-Based Image Retrieval), a technique for retrieving images on the basis of automatically-derived features such as color, texture and shape to aid diagnosis by identifying similar past cases. This report provides a comprehensive survey of the technical achievements in this research area. It clarifies some of the issues raised by this new technology, by reviewing its current capabilities and limitations, and its potential usefulness to users in medicine. The survey includes interesting papers covering the research aspects of medical image feature representation, multidimensional indexing, and system design. The report begins by discussing the operating conditions for content-based retrieval in the medical domain. Evaluation in content-based image retrieval is discussed next. Finally, based on the state of the art and the requirements for real-world applications, open research issues are identified and promising directions for future research are suggested.

Keywords: Medical images, Retrieval by content, Medical image database systems, PACS, HIS, MRI.

[1]The completion of this research was made possible thanks to Bell Canada's support through its Bell University Laboratories R&D program.

Contents

List of Figures

List of Tables

1 Introduction

Interest in the potential of digital images has increased enormously over the last few years, fuelled at least in part by the rapid growth of imaging in many domains such as the World Wide Web [73, 74, 76, 85, 86, 87] and medicine [10, 37, 52]. Users in many professional fields are exploiting the opportunities offered by the ability to access and manipulate remotely stored images in all kinds of existing new ways. However, they are also discovering that the process of locating a desired image within a large and varied collection can be a source of considerable frustration. Consequently, the problems of image retrieval are becoming widely recognized, and the search for solutions is an increasingly active area for research and development [8, 28]. With traditional text-based methods of image retrieval there are two major difficulties, especially when the size of image collections is large (tens, hundreds, or thousands). One is the vast amount of labor required for manual image annotation. The other difficulty, which is more fundamental, results from the rich content of the images and the subjectivity of human perception [65]. That is, the same image content may be perceived differently by different people. This perceptual subjectivity, coupled with imprecise annotation can cause irreparable mismatches in later retrieval processes. These problems have led to increased interest in techniques for retrieving images on the basis of automatically derived features such as color, texture, and shape. For example, in therapy treatment planning, the therapist is often interested in retrieving historical cases that exhibit a particular image feature. This research subject is now generally referred to as Content-Based Image Retrieval (CBIR) [26]. A number of survey articles have been published on CBIR [68, 67, 60, 82, 94, 1], but they deals with its general uses. However, the design of a CBIR system should take into account the needs of users and the specific nature of the data. Studying CBIR systems as a function of their context, their applications and the data they use gives us a better picture of CBIR research and generates numerous recommandations on how these systems can be adapted to users' needs. Kherfi et al. presented a thorough study of image retrieval on the Web [40]. In this report, we examine existing CBIR systems dedicated to the medical domain. These systems have several unique features, due to characteristics of medical images such as multiplicative noise, as we will explain in the next section. These characteristics require the use of imaging tools for color constancy,

spatial resolution, contour correction [17], etc. Moreover, content-based image retrieval from a database of medical images cannot be carried out using completely automated approaches; the presence of an expert is indispensable [78].

This report is organized as follows: The next section surveys medical image databases. To aid the reader in understanding the rest of the paper, Section 3 presents the problems underlying CBIR. In Section 4, we analyze methods for evaluating CBIR systems. To give an overview of CBIR in the medical domain, some existing medical image retrieval systems are examined in Section 5. In Section 6, we analyze the state of the art.

2 Background

In the health care domain, a huge amount of data is generated every day [3]. These data include alphanumeric structured data (e.g. demographic information), free text with imprecise medical terms and descriptions (e.g. pathology reports), images (e.g. computed tomography, magnetic resonance), and voice data. Fig. 1 shows various medical data repositories [16] which are intensively consulted using various retrieval systems such as Hospital Information Systems (HIS), Radiology Information Systems (RIS), Picture Archiving and Communication Systems (PACS), and (4) various research database systems (pathology, genome mapping, brain mapping, etc). Because there are many specialized branches of medicine, database systems are developed independently, their design reflecting the innovativeness of the database implementors, the scope of data required for their operation, and the culture of the particular department. The database system is only a first step in managing the available data. Some form of cataloguing and indexing is still necessary. Research involving data stored in multiple databases is often hindered by the fact that these databases function under different operating systems, with different data access and manipulation languages and different communication protocols. In addition, to provide the researcher with more data, scientific and medical databases require better query answering capabilities [16]. In this report, we will be interested in types of images which merit a particular attention because of their importance for physicians, the complex treatment they demand and their considerable number compared to other data generated. For example, a typical 700-bed teaching hospital conducts about 200,000 radiological studies per

year, generating over a million images [12]. The Georges Pompidou Hospital in Paris is a new middle-sized hospital with 827 beds; 360,000 imaging studies are expected to be performed each year [5]. We will also consider the various types of textual data which permit the utilization of these images. Alphanumeric data such as statistics and text, can be processed by many methods [41]. Finally, we should mention that voice analysis is not the subject of this report. Now, let us look in more detail at medical images and the needs of health care experts.

2.1 Characteristics of medical images

Before the acquired images can be used by information systems, these images must be digitized. However, the process of digitization does not in itself make image collections easier to manage. Medical images arising from pathology (endoscopy, histology, dermatology, etc), radiographic projection (X-rays, some nuclear medicine, etc), and tomography (CT, MRI, ultrasound, etc) impose unique, image-dependent restrictions on the nature of the features available for CBIR [91]. It is of foremost importance to consider whether the image arises from a projection technique such as conventional radiography or from a tomographic technique such as magnetic resonance imaging. We will present each of these in turn.

2.1.1 X-ray images

One of the earliest applications of X-rays was in medicine, where they were used for both diagnosis and therapy. Today, X-rays are still most widely used in this field [4]. They penetrate soft tissues but are stopped by bones, which absorb them. Thus if a photographic plate is placed behind a part of the body and an X-ray source is placed in front, X-ray exposure will result in a picture of the inner organs (Fig. 2). In the resulting image, dense tissues show up as light or white regions, while tissues that are easily penetrated by X-rays appear dark. X-ray images have the following characteristics:

- Distinguishing objects can be difficult. This is especially true when looking at multiple soft-tissue objects with low absorption characteristics that overlie one another. Structures may be obscured by overlying organs, or soft tissues may be insufficiently delineated for clear viewing.

- Multiplicative noise and granularity, may be present.
- The three-dimensional shape of objects such as internal organs can be hard to envisage.

2.1.2 Computed Tomography (CT) [38]

Conventional X-ray techniques have a major disadvantage: structures may be obscured by overlying organs. This issue is of particular importance in localizing brain tumors and other damaged sites in the brain. For such applications, a new form of X-ray process was developed, called computed tomography, or CT, formerly known as Computerized Axial Tomography, or CAT. In this process, the patient is placed inside an X-ray machine, and a narrow beam of X-rays sweeps across an area of the body, moving through a slight angle after each X-ray pulse (Fig. 3). The resulting series of X-ray images, taken from a different angles, appear with lighter brightnesses (Fig. 4). Computed Tomography is used in three primary modes. The original technique is the transmissive mode and uses an X-ray source. X-rays are transmitted through the imaged object and received at X-ray detection devices. The received signal at the detector is proportional to the density of the elements of the imaged object. The emissive mode of computed tomography relies on the emission of a detectable signal from the imaged object. The object can be directly excited or a substance can be introduced that is excited. In either case, detectors receive the emitted signals. The third mode of computed tomography is the reflective mode. As in the transmissive mode, a source transmits a signal directed at the imaged object. Instead of passing through the object, the signal enters the object and is reflected by the internal elements of the object back out to a detector device. The signal received at the detector is proportional to the density of the elements of the imaged object. Images produced by CT have the following characteristics:

- They are subject to multiplicative noise, because CT is in fact a form of X-ray process.
- They yield three-dimensional structures.

To produce high quality medical images, imaging techniques such as Magnetic Resonance Imaging (MRI) and Ultrasound imaging can be used.

2.1.3 Magnetic Resonance Imaging (MRI) [32]

MRI is an emissive mode imaging technique used primarily in medical settings to produce high quality images of the inside of the human body. It is based on the principles of Nuclear Magnetic Resonance (NMR), a spectroscopic technique used by scientists to obtain microscopic chemical and physical information about molecules. MRI started out as a tomographic imaging technique; that is, it produced an image of the NMR signal in a thin slice through the human body. After image acquisition is complete, co-registration, image enhacement and rendering are required. Images produced by MRI have the following characteristics:

- Contrast resolution is excellent.
- They yield three-dimensional information.
- Additive noise may be present.

2.1.4 Ultrasound imaging [34]

Ultrasound imaging (also called ultrasound scanning or sonography) is a relatively inexpensive, fast and radiation-free imaging modality. Ultrasound is excellent for non-invasively imaging and diagnosing a number of organs and conditions, without X-ray radiation. Ultrasound is a reflective mode technique which can show fetal development and bodily functions such as breathing, urination, and movement. Ultrasound is also extensively used for evaluating the kidneys, liver, pancreas, heart, and the blood vessels of the neck and abdomen. Ultrasound can also be used to guide fine-needle tissue biopsy to facilitate sampling of cells from an organ for lab testing (for example, to test for cancerous tissue). The ultrasound process involves placing a small device, called a transducer, against the skin of the patient near the region of interest. The ultrasound transducer combines functions like those of stereo loudspeaker and a microphone in one device: it can transmit and receive sound. This transducer produces a stream of inaudible, high-frequency sound waves which penetrate the body and bounce off the organs inside. The transducer detects sound waves as they bounce off or echo back from internal structures and organ contours. Different tissues reflect these sound waves differently, causing a signature which can be measured and transformed into an image. These waves are

received by the ultrasound machine and turned into live pictures via the use of computers and reconstruction software. Because high-frequency sound waves cannot penetrate bone or air, ultrasound is especially useful in imaging soft tissues and fluid-filled spaces. Images produced by ultrasound have the following characteristics:

- They are high-contrast.
- They yield three-dimensional information.
- Multiplicative noise may be present.

2.2 User needs and characteristics of medical queries

Medicine and related health professions use and store visual information in the form of X-rays, ultrasound or other scanned images, for diagnosis and monitoring purposes. There are strict rules on the confidentiality of such information. The images are kept with patients' health records which are, in the main, manual files, stored by unique identifier. Visual information derived from them, may be used for research and teaching purposes, for example, detecting and diagnosing lesions and tumors and tracking progress/growth by effective image processing (e.g. boundary/feature detection). What kinds of query are users likely to put to a medical database? To answer this question in depth, detailed knowledge of user needs (why users seek images, what use they make of them, and how they judge the utility of the images they retrieve) must be obtained. From the medical point of view, there are several applications [21] for automated content-based image retrieval:

- Automatic retrieval of relevant images may be useful for follow-up studies within a PACS [66].
- A medical student may have a set of images and wish to explore possible diagnoses.
- One may wish to display detailed data for analysis to locate other treated lesions in the database that are similar to the current case with respect to size, shape, intensity, and growth.

- A general practitioner may confirm her/his diagnosis of a specific patient and explore possible treatment plans by consultating the medical knowledge bank via the Web.

Accessing a desired image from a repository might thus involve a search for images depicting specific types of object or scene, evoking a particular mood, or simply containing a specific texture or pattern. Potentially, images have many types of attribute which could be used for retrieval, including the presence of a particular combination of color, texture or shape features; the presence of named events; the presence or arrangement of specific types of object or the depiction of a particular event. Each of the query types listed above represents a higher level of abstraction than its predecessor, and each is more difficult to answer without reference to some body of external knowledge. This leads naturally to a classification of query types into three levels of increasing complexity:

- Level 1: Comprises retrieval by primitive features such as color, texture, shape or the spatial location of image elements. This level of retrieval uses features which are directly derivable from the images themselves.

- Level 2: Comprises retrieval by derived features, involving some degree of inference about the identity of the objects depicted in the image (e.g. *find a picture of a coronary artery.*)

- Level 3: Comprises retrieval by abstract attributes, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted (e.g. *find an image of a patient who has suffered a blockage of the right coronary artery.*)

3 Content-based image retrieval

Content-based image retrieval provides an alternative approach to the traditional information retrieval paradigm [68, 82, 96, 97]. It is the process of retrieving desired images from a large collection on the basis of features (such as color, texture and shape) that can be automatically extracted from the images themselves. Research and development issues in CBIR cover a range of topics [60, 67, 63, 50, 85, 9], many of them shared with mainstream image processing and information retrieval. Some of the most important are:

- Understanding image users' needs and information-seeking behavior.
- Identifying of suitable ways of describing image content.
- Extracting such features from raw images.
- Providing compact storage for large image databases.
- Matching query and stored images in a way that reflects human similarity judgements.
- Efficiently accessing stored images by content.
- Providing usable human interfaces with CBIR systems.

Thus, the issues which must be addressed by a retrieval system which supports queries based on image content may be summarized as follows [29]:

- Retrieval methods based on similarity, as opposed to exact matching.
- Selection, derivation, and computation of image features and objects that provide useful query expressiveness.
- A user interface that supports the visual expression of queries and allows query refinement and navigation of results.
- Indexing of the database in a way that compatible with the expressiveness of the queries.
- Taking into account distributed databases as well as the fusion of images produced by different sensors and thus providing different information.

Many image retrieval systems can be conceptually described by the framework depicted in Fig. 5. At the heart of this framework is a database structured with respect to image features which are extracted for both data entry and query interpretation and compared for similarity during retrieval. The user interface supports the *closing of the loop* which relates the formulation of queries to the browsing, to retrieve images, or to summarize the database. We will now present the main problems underlying CBIR.

Color	Histograms, color co-occurrence histograms.
Shape	Moments, template matching, size functions, edges.
Texture	Directionality, periodicity, randomness, Covariance.
Other	Wavelet coefficients, invariant features

Table 1: Overview of commonly used features in CBIR

3.1 Feature extraction and metadata production

Feature extraction is the basis of Content-Based Image Retrieval (See Table 1). In a broad sense, features may include both text-based features (keywords, annotations, etc) and visual features (colors, textures, shapes, faces, etc). Within the visual feature category, the features can be further classified as general (e.g. color, texture, shape) or domain-specific (e.g. geometric and anatomical location [43, 44]). The latter group is application-dependent. Because of perceptual subjectivity, there is no single best presentation for a given feature. In fact, for any given feature there are multiple representations which characterize it. For example, some representative studies of color perception and color spaces can be found in [47, 95], and a summary of the different features is given in [67]. We can classify features into two groups:

- Low-level features: color [33], color layout, shape, texture [64, 84], etc.
- High-level features: spatial and temporal relations, object's name, etc.

We will now present one of the most important high-level features used in CBIR, metadata. Metadata is data about the data or information that describes the data. Sometimes, the two concepts (data and metadata) may be confused. In fact, the distinction between data and metadata is not always clear. The emerging interest of researchers in the use of metadata is motivated by two main factors: on one hand, metadata use has been investigated as a way of managing complex multimedia documents efficiently [20]. On the other hand, the use of metadata seems promising in order to solve problems in heterogeneous systems. Modern health care systems present both of the features discussed above: diagnostic tests produce a huge amount of data from different media, and health care organizations are characterized by distributed structures which manage their own information in heterogeneous environment [10]. For an

efficient use of diagnostic data, it is important for physicians to be able to access multimedia data by content. Moreover, effective solutions are needed in order to share information about patients whose health care history involves other health organizations. A medical report can be modeled as a hypermedia structured document which consists of five main components: general data about the patient, data about the test, multimedia data (e.g. images), semiotic description of multimedia data, and a diagnostic conclusion, which may be produced by a radiologist [14]. To take another example, a wide variety of images may also be acquired by an ophthalmologist in clinical practice. These images must be integrated into the medical record and must be available for archiving, searching, and retrieval. An ophthalmic image can be annotated with literally hundreds of tags, recording image properties and their relationships (e.g. patient data, administrative data, prior medical history, ocular history, laboratory results, etc). It is particularly important that these data be machine-readable and Web-accessible. Use of the Web reduces proprietary hardware and software barriers and facilitates the exchange of information [56]. But, the growth of a global information network, mainly supported by the success of Internet and WWW, is emphasizing the development of systems characterized by both of the problems discussed above (complex multimedia documents, heterogeneous systems). For the first of these, metadata becomes important as it makes it possible to perform content-based search on not-traditional data such as images, video, and so on. For the latter, metadata is used to browse, navigate and retrieve information with the focusing on information content, without relying on the structure and organization of the individual databases. A very important language used in the medical domain, in which highly specialized concepts and attributes exist and new ones are regularly created, is XML (eXtensible Markup Language) [23], which allows data description tags to be extended and customized according to the domain user's specifications. In contrast to HTML (HyperText Markup Language) [30], which is solely concerned with display, XML adds semantic information to the markup so that document searching and retrieval is much easier. XML imposes structure on data and allows a logical tree to be built that represents data elements and attributes. The advantages of encoding medical and image data using XML are manifold. The benefits of structured data in medical care include vastly improved workflow and content management. It also enables measurement of outcomes, foster quality assurance, and improves medicolegal documentation. Unlike other standards, XML

supports the document-centered approach to medical records. (See Fig. 6).

3.2 Classification and indexing schemes

A major problem in dealing with large image database is efficiency of retrieval. One of the key issues in achieving such efficiency is the design of a suitable indexing scheme [2, 22]. The goal of indexing is to create a compact summary of the database contents to provide an efficient mechanism for retrieval of the data. To make content-based image retrieval truly scalable to large image collections, efficient hierarchical multi-dimensional indexing techniques need to be explored [98, 99, 90, 35]. For example, in medicine, there are many methods for classifying images. The inherent categorization of pathologies in medicine provides a way to classify medical images (cases). This content-based organization of medical images naturally forms a hierarchical structure in wich the bottom leaves correspond to a set of specific images (cases), and higher nodes correspond to a subcategory of pathological cases. Using this structure, cases are first stored by their pathological nature (tumor, bleeding, stroke, etc), then stored by their proximate anatomical locations (intra-axial, extra-axial), and lastly by their visual features. Microscopic histology or the macrophotography used in dermatology could suggest different approach to database organization because of their inherent properties of color, shading, and resolution. These last two examples suggest a database indexing scheme that could take into account visual features. Ultrasound images of large organs with relatively uniform tissue, such as the spleen or liver present relatively homogenous image patterns. These might be indexable by texture. In general, the existing popular multidimensional indexing techniques include k-d tree, priority k-d tree [99], quad-tree, K-D-B tree, hB-tree, R-tree and its variants R^+-tree and R^*-tree [31]. Chang [7] has proposed the use of 2D-strings for the indexing of images based on spatial relationships and attributes of the objects in the image. His methods have been extended by other researchers [62, 42]. In addition to the above approaches, clustering and neural nets are also promising indexing techniques [24]. For more information, very good reviews of various indexing techniques in image retrieval can be found in [98, 49].

3.3 Similarity and matching

When a user submits a query, the system searches the database containing features that represent the medical database images in order to retrieve all images that correspond to the query. If images are indexed, parsing time will be reduced because the system will no longer be forced to search the whole database but only the index. A first question to be answered is whether the system uses matching or similarity to find the designed images. In traditional databases, matching is the fundamental operation, it consists of comparing an item with the query, and deciding whether, or not the item satisfies the query [61]. In textual databases, matching is a binary operation, in other words it involves a decision whether the item matches the query or not. In similarity-based search, on the other hand, images are ordered with respect to similarity with the query, given a fixed similarity criterion [72]. The results consist of the images most similar to the query. Image databases should rely on similarity rather than on matching [71]. In fact, one way of assigning a meaning to an observed feature set is to compare a pair of observations using a similarity function. Requiring robust matching is not a satisfactory solution because there may be a slight difference between a database image and the query due to some accident, imperfection, geometric transformation, or change in illumination; or the two compared images may be similar but not necessarily identical. In searching for a query image among the elements of the data set of images, knowledge of the domain will be expressed by formulating a measure of the similarity between images on the basis of some feature set. It should also be noted that there is another view in which similarity is seen as an essentially probabilistic concept. This view is rooted in the psychological literature [59] and the context of content-based retrieval, where the notion of similarity used should be as close as possible to human similarity since human judgement is the reference. Similarity perception is a complicated activity. It results from the cooperation of a number of different mechanisms placed at different levels in the visual system. Because of this, it is difficult to give a unique characterisation of similarity perception. A number of similarity models have been proposed [71]. Early models [92] hypothesized that similarity assessment was based on the measurement of a suitable distance in a psychological space. Another class of similarity models, strongly connected to the metric approach, is the *Thurstone Shepard class*, based on an idea that goes back to [92]; these are reviewed in [25]. In such models, the similarity between two images is a function of the Minkowski distance given

by:

$$d(x,y) = [\sum_{k=1}^{n} |x_k - y_k|^{\gamma}]^{\frac{1}{\gamma}} \quad (1)$$

In more recent years, some models that abandon the strict distance model have been developed. Among these models we can find the work of Amos Tversky [93], who proposed the *feature contrast model*. Instead of considering images as points in a metric space, he characterized them as sets of features. Let a, b be two images and A, B the respective sets of features. The results implies that the similarity can be obtained using a linear combination (contrast) of a function of the common features ($A \cap B$) and the distinctive features ($A - B$ and $B - A$). Mathematically, this can be written:

$$S(a,b) = f(A \cap B) - \alpha(A - B) - \beta(B - A) \quad (2)$$

where f is a non-negative function, α and β are two constants and S is the similarity. Another problem for similarity based searches is how to deal with complex queries that include operations like ordering with respect to two or more similarity measures (*show me images with this dominant color and this texture*), or with respect to two or more images (*what is there similar to either of these images?*). Complex queries can be constructed from simple queries using the connectors AND, OR and NOT. The similarity corresponding to a complex query can be defined as a function of the similarities corresponding to its components simple queries. Similarity can be computed by measuring the resemblance or the difference between two images and is based on the features describing the two images [71]. It is a subjective concept that must consider the image features, the application, the specific needs of the users and his attachment to features. For more information about similarity, a very good review is given in [82].

3.4 Relevance feedback

Since subjectivity and imprecision are usually associated with the specification and interpretation of subjective attributes, the query processor should be designed to deal interactively with these problems at the time of query specification or processing. The query interface can be designed to guide users through the query-specification process and facilitate user-relevance feedback [75]. Relevance feedback can be seen as an interactive process in which the user judges

the quality of the retrieval performed by the system by marking, among the images retrieved by the system, the ones he/she perceives as truly relevant. This information is then used to refine the original query, resubmitted for a sharper selection. User involvement in providing the relevance feedback should be at a conceptual level. That is, users should not be forced to explain feedback in terms of low-level image features. For example, the user may provide relevance feedback by simply labeling a set of retrieval images relevant, nonrelevant, or somewhat relevant. The user should also have the possibility of indicating the class of features (e.g. shape, texture) relevant for retrieval, and specifying the weight of each feature in the retrieval process. Relevance feedback has been shown to be a very effective tool for enhancing results in text retrieval [89]. In CBIR it is more and more frequently used and good results have been obtained [70, 100, 69, 39]. The two main strategies for relevance feedback are [6]:

- Making separate queries for each feedback image and merging the query results, or
- Creating a *pseudo-image* from the feedback images and executing a query with this image.

Two types of feedback are possible:

- Positive feedback: It is limited to preselected images and weights the features of these images more strongly. All high-ranked returned images have many features in common.
- Negative feedback: It can greatly improve the query result, but it is important to use the right images as negative feedback so as not to inhibit any important features.

Different methods have been proposed for implementing relevance feedback. In some systems, like I^2C [52, 51, 58] or the one proposed by the Robotics Institute [43, 46, 45], the user can assign weights to features to improve the feedback. In the IGDS [17, 18, 19] system, the user is asked to introduce voice and graphical input at the begining, after which visual and audio feedback are possible. A very interesting approach, based on the use of negative examples, is proposed by Kherfi et al. [39].

4 Evaluation in content-based image retrieval

Evaluation of retrieval performance is a crucial problem in content-based image retrieval. Different methods for measuring the performance of a system have been created and used by researchers [83]. Most of the measures used in CBIR have long been used in text retrieval [88]. Two examples of such measures are:

$$precision = \frac{\text{No. Relevant documents retrieved}}{\text{Total No. documents retrieved}} \tag{3}$$

$$recall = \frac{\text{No. Relevant documents retrieved}}{\text{Total No. Relevant documents}} \tag{4}$$

A good information retrieval system should maximize these two measures. We note that these measures require ground truth and can't be used in this form to evaluate a medical image retrieval system, especially one that is Web-oriented, since neither the total number of images on the Web nor the number of images relevant to a given query can be known. However, we can consider other evaluation measures:

- Relevance and accuracy: Ground truth is replaced by the human subjectivity; i.e., either the system returns the results the user expects or not. We can distinguish different degrees of relevance, ranging from retrieving exactly the images the user is looking for, through retrieving images that are close to the query or relevant to the same subject but not exactly what the user is looking for, to retrieving images that are totally irrelevant to the query.

- Retrieval time: After the user submits his query, the system looks in its index and performs comparisons between images. This operation must be as quick as possible in order to return the results to the user in an acceptable time.

- Ease of use: The visible part of the system, its interface, must be easy to use in all stages: formulating queries, selecting sample images and specifying regions of interest in these images, and displaying retrieval results.

As the user is the judge, these measures can be estimated by asking a number of representative people to use the system and evaluate its performance by giving a grade indicating their degree of satisfaction.

5 State of the art

In this section, we give details on the following systems: KMeD, I^2C, I^2Cnet, a system proposed by *The Robotics Institute*, IGDS, and ASSERT. Each system is analyzed on the basis of the following features: objectives, queries, retrieval methods, relevance feedback, user interface, and feature extraction.

5.1 KMeD: Knowledge-Based Multimedia Medical Distributed Databases [10, 13, 15, 11]

Developer: Department of Computer Science, University of California, Los Angeles.
Homepage: http://www.kmed.cs.ucla.edu/
This system utilizes a knowledge-based approach to retrieve medical images using spatial and temporal constructs. Selected medical images (e.g. X-ray, MRI) are segmented, and contours are generated from them. Features (e.g. shape, size, texture) and relations (e.g. spatial relationships among objects) are extracted and stored.
Objectives:

- Query medical distributed databases by both image and alphanumeric content.
- Model the temporal, evolutionary and spatial natures of objects (e.g. bone growth) and enable queries based on this modeling.
- Formulate and answer conceptual and imprecise queries by relaxation, and provide relaxation control to satisfy user constraints.
- Provide relevant *value-added* information as part of the query answer even though it is not explicitly requested (an ability called associative query answering).
- Provide a domain-independent high-level query language and a medical domain-oriented, graphically interactive user interface.
- Provide analysis and presentation methods for display of data and knowledge models.

Centralized/Distributed system: Distributed databases.

Data: Computed Tomography, MRI (Sagittal Magnetic Resonance Images), HIS information (patient's sex, date of birth, annotations,etc), X-rays and sounds.

Indexing: Use of data structures such as R-trees and B-trees.

Queries:

- Example 1: Temporal object management using a predicate to restrict an evolutionary event. *Retrieve an image sequence of a patient demonstrating the fusion of the thumb metacarpal.*
- Example 2: Query by image content and cooperative query processing using abstraction hierarchies with association. *Retrieve all hand X-rays of 12 year old Korean American patients with Turner's Syndrome.*
- Example 3: Spatial object management. *Retrieve all image cases demonstrating the invasion of an adenoma into the sphenoid sinus in pre-adolescent patients.*

Retrieval methods: KMeD includes a cooperative query answering layer which uses domain knowledge and inference techniques to make intelligent decisions on localizing a query's search space, regulating a query's solution space, conceptualizing complex data entities and processes, and associating context-dependent subjects.

User/Feedback:

- Graphical user interface employing both textual and graphical queries.
- Ability to control the size of returned answers.

Features:

- Low-level: Region (segmentation based on wavelet transforms,etc), edge (detection requiring user interaction), contour, intensity discrimination, bounding box, area, volume diameter, length, circumference, text.
- High-level:
  - Spatial relations: Spatial relations between a pair of objects include:

* Orthogonal relations: Describe directional relationships between objects, such as East, South, and Southeast.
* Containment relations: Describe the relative position and the locations of contact between a pair of objects, such as invades and contains.

– Temporal relations.

5.2 I^2C: Image Indexing by Content [52, 51, 58]

Developer: Institute of Computer Science, Foundation for Research and Technology-Hellas and Department of Computer Sciences, University of Crete, Heraklion, Crete, Greece.

Homepage: http://terpsi.ics.forth.gr/ICS/acti/cmi_hta/activities/software/i2c/i2c.html

I^2C is an information system for the indexing, storage, and retrieval of medical images by pictorial content. As illustrated in Fig. 7 it is a modular extensible system which has been developed based on object-oriented principles. The I^2C architecture incorporates a set of tools and algorithms for the extraction, indexing, and storage of image descriptions. These include noise reduction, segmentation, and line approximation algorithms, as well as a contour editor and a storage manager. Some of these tools are interactive and others are automatic. The system is open in the sense that new tools may be added to the system with minimal effort, in some cases without even disrupting its operation. I^2C provides an integrated environment for the definition and management of content descriptions of medical images and appropriate similarity criteria, which capture and possibly extend the knowledge of experts, so that medical images are compared not only on the basis of their morphological characteristics, but also their clinical content. The basic features of this system are:

- indexing and retrieval of medical images by pictorial content;
- an image browser;
- image processing and analysis tools;
- interactive construction of pictorial content descriptions.

Objectives:

- Image indexing, storage and retrieval.
- Browsing.

Centralized/Distributed system: Distributed databases.

Data: MRI.

Indexing:

- Hash, B-tree indices or structures combining these two types of indices.
- Individual medical images are identified with a byte string called imageID.

Queries: Sketch or image queries.

Retrieval methods: Description types are used in I^2C to encapsulate information relevant to an image content description method and a content-based retrieval strategy. The structural components of a description type are: the description generator, the description manager, and the description matcher. The description generator produces the logical image, which consists of a set of persistent objects. The description manager manages logical images in the logical database of the description type. Finally, the description matcher processes content-based queries addressed to the description type and identifies images similar to the query image. The imageIDs of the images that constitute the response to the query are reported back to the system, the raw image data are retrieved, and their miniature are displayed on the image browser.

User/Feedback:

- The users are allowed to search via relations among objects, relative position, texture, orientation, border type.
- Presence of a browser and of segmentation and contour tools.
- Insertion of annotations to the images, setting of weights.
- Display of results of image queries (up to 18 simultaneous pictures).

Features:

- Low-level: Area, texture, border type, region, contour.
- High-level: None

5.3 I^2Cnet: Image Indexing by Content network [56, 53, 54, 57, 55]

Developer: Institute of Computer Science, Foundation for Research and Technology-Hellas and Department of Computer Sciences, University of Crete, Heraklion, Crete, Greece.

*Homepage:*http://terpsi.ics.forth.gr/ICS/acti/cmi_hta/activities/software/i2cnet/i2cnet.html.

I^2Cnet extends the functionality of I^2C which can serve as a browser for images and image descriptions, an editor of image content descriptions, and a processor of content-based queries. I^2Cnet addresses these issues by providing content-based retrieval as an added-value service in a regional health care network. The main elements of the I^2Cnet architecture are: I^2C clients, servers, and brokers. I^2C clients use a standard WWW browser to request I^2C services and submit content-based similarity queries. I^2C service brokers activate software agents to update the profile of available services and provide support for network-transparent queries. I^2C servers maintain databases of image content descriptions and interact with the health care network to retrieve additional information on selected images and respond to queries which involve image content and other electronic patient record data. A major advantage of this approach is that other services like those of QBIC [27], can be accessible through a unified framework offered by HTML. The combination of standard HTML and browser programming offers a generalized interface to heterogeneous types of data, and provides advanced, platform-independent, client-server interaction.

Objectives: The representation, storage, and retrieval of medical images based on different descriptions of the image.

Centralized/Distributed system: Distributed databases.

Data: MRI.

Indexing: The same methods of indexing used by I^2C.

Queries: Sketch or image queries.

Retrieval methods: A matching algorithm retrieves from the database images whose descriptions match the query description under the specific similarity criterion.

User/Feedback:

- The combination of standard HTML and browser programming offers a generalized interface to heterogeneous type of data, and provides advanced, platform-independent, client-server interaction.
- Network-transparent interaction, in which the user formulates a request without any concern as to which server or servers will process it.
- Server-specific interaction, in which the user specifies how the request should be handled.
- The user interface of an I^2C client is a typical Web browser like Mosaic, Netscape, etc. The user submits content-based queries by interacting with Web pages (effectively filling out HTML forms), and then browses the retrieved images.

Features:

- Low-level: Area, texture, border type, region, contour.
- High-level: None.

5.4 System proposed by *The Robotics Institute*, Carnegie Mellon University [43, 46, 45]

Developer: The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Homepage: http://www.ri.cmu.edu/projects/project_281.html

This is an image retrieval framework centered around classification-driven search for a weighted similarity metric for image retrieval. It uses an approach rooted in Bayes decision theory. In summary, the approach uses memory-based learning to explore, discover and manipulate image feature space, in the hope of finding the most direct, effective and economical mapping from a proper subset of non-uniformly rescaled image features to their corresponding image classes. This image retrieval framework consists of three stages, as shown in Figure 8: (1) feature extraction; (2) feature selection via image classification ; and (3) image retrieval.

Objectives:

- Construct creative statistical image features such that the image semantics are captured with high probabilities.
- Use the most discriminative feature-subset as the front-end index to find (for image classification or retrieval) medically similar cases in a large image database to aid diagnosis, surgical planning, patient treatment, outcome evaluation and medical education.

Centralized/Distributed system: Distributed databases.
Data: CT, MRI and HIS (patient's age, sex, symptoms, test results, etc).
Indexing: A memory based learning (MBL) technique, called *Kernel regression*, is applied to classify images.
Queries: An entire medical image or a part of it.
Retrieval methods: Using k-nearest neighbors in the selected feature space: the weighted subset of the features that yield the best performance in classification is used as a similarity metric for image retrieval.
User/Feedback: The user assigns the weights of features used for the retrieval.
Features:

- Low-level: Size, contrast, boundaries, density, shape, global statistical properties, mean of gray-level intensity, deviation of gray-level intensity, and local regional statistics.
- High-level: Geometric and anatomic location.

5.5 IGDS: Image Guided Decision Support System for Pathology [17, 18, 19]

Developer: Department of Electrical and Computer Engineering, Rutgers University, Piscataway, USA.
Homepage: http://www.caip.rutgers.edu/~comanici/jretrieval.html
This system has been developed for multimodal indexing, querying, and retrieval of medical information from consensus-graded archives of digitized images. The system allows physicians to interactively review diagnostic images and to delineate regions containing structures which are either unidentifiable or are known to be key to the diagnosis. The architecture of this

system is given in Fig. 9.

Objectives:

- Locate, retrieve and display cases which exhibit morphological profiles consistent with the case in question and assist pathologists to discriminate wrong malignant lymphomas and microscopic specimens.
- Help physicians and technicians during routine screening and analysis.

Centralized/Distributed system: Distributed databases.

Data: Images corresponding to malignant lymphoma, leukemia and benign cases: 450×350 true color pixels taken with a high-resolution video camera.

Indexing: The database indexing is performed off line. A module is used for the analysis and registration of the incoming cases. Then, the weights of the dissimilarity measure are re-learned to account for the new entries in the database.

Queries: Load a query image and select a region of interest (ROI).

Retrieval methods: Dissimilarity metric defined as a linear combination of the normalized distance corresponding to each visual attribute. The retrieval process is multithread (simultaneous access to the database being permitted).

User/Feedback: Retrieval can be browsed using voice or graphical input. Audio feedback is provided by speech analysis. The users can modify the color and spatial resolutions (for experiments and maintenance). The system provides a user-handled contour correction tool based on cubic splines. Users can select different query attributes, browse the retrieval, select a different scale for viewing and display specific clinical data and video clips. Input commands can be formulated by voice or graphical input.

Features:

- Low-level: Shape, area, texture, color, contour, region, texture and text. Based on multiresolution simultaneous autoregressive model.
- High-level: None.

5.6 ASSERT: Automatic Search and Selection Engine with Retrieval Tools [37, 79, 80, 78, 81, 77]

Developer: The Department of Radiology at Indiana University, the School of Medicine at the University of Wisconsin and the Machine Learning Lab at Purdue University.

Homepage: http://rvl2.ecn.purdue.edu/~cbirdev/WWW/CBIRmain.html.

This is a human-in-the-loop CBIR system for medical images. A unique feature of this system is that it is interactive. In fact, it seeks a physician's help in areas that cannot be fully automated like the segmentation of pathologies. An important characteristic of this system is that it uses local features, not global ones like the majority of CBIR systems.

Objectives: Provide an opportunity to aid physicians in the process of diagnosis.

Centralized/Distributed system: Centralized databases.

Data: HRCT (High Resolution Computed Tomography) lung images.

Indexing: An efficient algorithm called *Multi-Attributed Hash Table Indexing* is used to archive and index the medical image database. Based on the pre-selected attribute set, the system creates a decision tree and then translates this decision tree into a multi-attribute hash table. An index is computed from the hash table for archiving the images.

Queries: An entire medical image or a part of it.

Retrieval methods: Observing an abnormality in a diagnostic image, the physician can query a database of known cases to retrieve images (and associated textual information) that contain regions with features similar to what is in the image of interest. Thus, a first classification of the query image under a disease is done according to a set of features. The images the most similar to the query image are retrieved using the same set of features. Similarity is defined by Euclidean distance.

User/Feedback: User intervention is necessary because the Pathology Bearing Regions (PBRs) in the images used cannot be segmented out by any of the state of the art segmentation routines due to the fact that for many diseases, these regions often do not possess sharp edges and contours. Thanks to the graphical interface, it takes a physician only a few seconds to delineate the PBRs and any relevant anatomical landmarks. A benefit of this approach is that when a query image contains more than one pathology, the physician can choose to circumscribe only

one of the regions in order to focus retrieval on that pathology. The system permits feedback, also.

Features:

- Low-level: Region, shape, texture, and edges.
- High-level: None.

6 Critical analysis

After surveying most of the existing systems, we will now give a brief critical analysis of them, based on the points evoked in Section 3, that is, the consideration of user needs, relevance feedback, features used, similarity, and indexing.

6.1 Consideration of user's needs

User needs are supported in different manners by the above systems. If we consider the possible queries, we note systems like I^2C which allow sketch queries; systems which allow an entire medical image or a part of it as a query such as ASSERT; and systems which enable the use of predicates to restrict evolutionary events and provide more useful responses to a given query, like KMeD. Users may sometimes be interested in the content of the image (the presence of a given shape). Such queries are relatively easy to express and there are systems which allow this such as I^2Cnet. However, we can't find a system which combines all of these types of queries. ASSERT, for example, is dedicated to high-resolution computed tomography lung images. Among the specialized systems we find IGDS, which is dedicated to digitized images corresponding to malignant lymphoma and leukemia. Another specialized system is I^2C which is dedicated to MRI. Another important point that must be taken into account is the issue of iconic queries which are user-generated. These may be sketches of features that are important or they may be prototypes. The use of icons and associations with prototypes will provide the user with a means of developing customized semantics. Generic schemas will be needed to provide a starting point for the schema developments so that a user can define which relations and which similarity measures are appropriate for the problem under consideration. A variety

of objects at different levels of abstraction will be required by users to support iconic queries and customized schemas. A pictorial-based query language will be essential for full utilization of a medical imaging database [36]. In certain systems such as I^2Cnet, the user can interact with the generation of the image description. First, the query image is segmented, a number of key ROIs are selected, and their features are computed. In fact, the majority of the systems allow the user to introduce queries based on a region of interest (ROI). The clinical researcher will also require tools that allow for end-user designed customized schemas for retrieval and search that can be edited, modified, and adapted to new queries. Queries to an image database by different users may make vastly different demands on the query language [12]. For example, a medical oncologist may want to generate complex queries about an image that relate to the functionality and/or structure of organs in the image. A database to be used for teaching, on the other hand, may require a means of accessing images that all exhibit a particular morphological characteristic. Thus, all of the systems described must not only embed the retrieval functionality, but also offer the possibility of categorizing images by general or specific subject. In fact, a catalogue or a resume of the database is very useful for a user who doesn't have a clear idea in mind about the image he is looking for. It allows him to browse through different subjects [97]. Some of the systems studied allow browsing (IGDS, I^2C), but none of them permit the use of negative examples. This functionality is very important because it enables the user to refine the query, so we think that its introduction in future systems would be really advantageous. In addition, it is clear that medical images represent a particularly unique class of problem for database design. In fact, the database schema must evolve considerably over the lifetime of the database. The changeability of the schema seems to be the single most important aspect of medical image databases, and much design effort must be focused on managing this change. Users must be able to generate queries of a set of medical images that are changing and dynamic [12]. A patient who undergoes a CT scan to delineate a primary tumor of the lung may subsequently be discovered to have liver metastases. Thus, the physicians accessing the database will want to incorporate new knowledge about liver metastases into the database and have the capability of developing relevant questions about the patient's condition, both in the past and in the present. As the user develops possible hypotheses for exploring a database, having the opportunity to navigate through the database collecting images that are

interesting will suggest new formalizations [21]. Thus, tools that allow for "*show me one like this but larger*" or "*show me one between these two*" or "*show me one that is very different*" may provide the user with powerful means of developing new conceptualizations and knowledge. This *changing the field of view* approach is perceived as an important attribute of a medical imaging database. Having a descriptive language is very important too, because the description of image features will lead to new knowledge and new categories for describing disease. Finally, to respond better to the different kinds of queries, any system must support relevance feedback. This question will be discussed in the next subsection.

6.2 Metadata and features used

Only some systems use metadata. The most significant example is KMeD, which uses a Temporal and Spatial Evolutionary Data Model (TEDM & SEDM), a unified data model for representing multimedia, timeline, and simulation data [12]. In our opinion, this model could be used in other medical image retrieval systems because it has proved its efficiency in KMED. However, this model could be improved by taking into account the notion of the Region Of Interest, which we believe is a very important aspect, especially in medical images. Thus, serious study must be devoted to producing a data model which supports all aspects of an image produced in the health care domain [48]. The majority of the systems described here use low and high-level features to index images. Table 2 summarize the features used by the different medical image retrieval systems that we have studied. If a system can make use of other data types (such as sound and video) it will be much better. However, using a large number of features can cause certain difficulties such as the problem of feature dimension which makes comparison difficult and renders the existing indexing methods unusable. Selection of pertinent features and relevance feedback can be good solutions. Concerning relevance feedback, most of the systems are based on it. We think that relevance feedback is important because one-pass retrieval often doesn't yield the desired results. Among general image retrieval systems, there are those which allow relaxation control by the user, the control of the size of returned answers, or a combination of the two. The majority of the systems we presented in the last section have a graphical interface which enables the user to interact with the system. Some

	KMeD	I^2C	I^2Cnet	*The Robotics Institute*	IGDS	ASSERT
Color	×			×	×	
Texture		×	×		×	×
Region	×	×	×		×	×
Contour	×	×	×		×	
Shape	×			×		×
Text	×				×	
Statistical properties				×		
Spatial Relations	×					
Temporal Relations	×					
Geometric and anatomic location				×		

Table 2: Features used by existing systems

systems, like the one proposed by *The Robotics Institute*, allow the user to adjust the weights of various visual features interactively [43, 46, 45]. This last technique seems to be the best way of retrieving the most relevant images because it satisfies the user needs. In fact, thanks to interaction with the system, the characteristics can be selected and the similarity computed. In our opinion, proposing a feedback method which allows this is very important. In this regard, the introduction of negative examples [39] becomes a real necessity.

6.3 Indexing and similarity

When we look at data indexing, some of the reviewed systems use it and some not. Using indexing can help to minimize retrieval time because the system is not required to search the whole database of features when it looks for images [90]. However, when dealing with systems which utilize generic databases like the Web (I^2Cnet for example), indexing techniques are not very effective. We think that the use of mixtures as an indexing technique is very efficient because of its simplicity [21] compared to data structures. Given an image, we compute posterior and prior probabilities for each image class, as this information is needed for further

processing. Another problem this raises is which features will be extracted from the images and used to construct the index. This problem is closely related to the notion of similarity, because any index should minimize the time required to find similar images, so the index schema should take into account the similarity measures that are used [78]. When developing a user-centered approach, the system designer needs to consider user requirements, and in particular, to know what the user means by *similar* images. Existing systems use different similarity measures, such as quadratic distance between color histograms and shape matching. These measures do not always correspond to human perception. Although relevance feedback can help, we think more studies should be done to define similarity measures which correspond better to human judgement. Similarity is a subjective notion and can have different meaning to different users, or even for the same user at different times. Two images that are considered very similar by one user may be considered very different by another, depending on the features on which the judgement is based. A possible solution for this subjectivity is the similarity learning, which is based on relevance feedback: after a user introduces his query, the system will have to analyze it and try to find out the most important features that are of interest to that user, weighting features on this basis. Note that similarity and indexing are related problems. This is very clear in the ASSERT system, for example. In this system, retrieval of the images most similar to the query image is facilitated by the way in which images are indexed. In our opinion, therefore, indexing must take retrieval methods into account.

7 Conclusion

In this report, we have discussed medical image retrieval, giving some examples of medical image retrieval systems. An interesting question regarding the topic of this survey is whether medical databases present novel problems to the content-based image retrieval community. In many ways, medical images are not similar to optical or radar images. However, like other databases, the goal of medical imaging databases is to provide a mean for organizing large collections of heterogeneous, changing, pictorial, and symbolic data. This must reside in a structured environment that can be synthesized, classified, and presented in an organized and efficient manner to facilitate optimal decision making in a health care environment. We believe

that responding better to user queries requires the system designer to understand user needs, which is why through study is needed on the actual use of medical image databases. Often, however, the user doesn't have a clear idea in mind about the image he is looking for. Thus, a catalogue or resume of the database is useful, allowing him to browse through different subjects. This will be more interesting if the system allow both positive and negative feedback to refine queries.

References

[1] Aigrain, P., Zhang, H. and Petkovic, D. Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review. *Multimedia Tools and Applications*, 3(3):179–202, 1996.

[2] Ardizzone, E. and Cascia, M. L. Content-Based Indexing of Images and Video Databases by Global and Shape Features. In *International Conference on Pattern Recognition (ICPR)*, pages 140–144, August 1996.

[3] Arya, M., Cody, W. and Faoutsos, C. QBISM: A Prototype 3-D Medical Image Database System. *Data Engineering Bulletin*, 16(1):38–42, 1993.

[4] Baxes, G. A. *Digital Image Processing: Principles and Applications.* John Wiley & Sons, Inc, 1994.

[5] Bozec, H. L., Zapletal, E., Jaulent, M. and Heudes, D. Towards Content-Based Image Retrieval in a HIS-Integrated PACS. In *AMIA 2000*, 2000.

[6] Celentano, A. and Chiereghin, S. Multiple Strategies for Relevance Feedback in Image Retrieval. Technical Report CS-99-8, Foscari University, Venice, 1999.

[7] Chang, S., Aho, A., McKeown, K., Radev, D., Smith, J. and Zaman, K. Columbia Digital News System, An Environment for Briefing and Search over Multimedia Information. In *Proceedings of (IEEE) International Conference on the Advances in Digital Libraries*, pages 82–95, Washington, DC, 1997.

[8] Chang, S. K., Shi, Q. Y. and Yan, C. W. Iconic Indexing by 2D Strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:413–428, 1987.

[9] Chang, S., Smith, J. R., Meng, H. J., Wang, H. and Zhong, D. Finding Images/Video in Large Archives. *D-Lib Magazine*, 1997.

[10] Chu, W. W., Cardenas, A. F. and Taira, R. K. KMED: A Knowledge-Based Multimedia Medical Distributed Database System. *Information Systems*, 20(2):75–96, 1995.

[11] Chu, W. W., Dionisio, J. D. N., Cardenas, A. F., Taira, R. K., Aberle, D. R., McNitt-Gray, M. F., Goldin, J. and Lufkin, R. B. A Unified Timeline Model and User Interface for Multimedia Medical Databases. *Knowledge and Data Engineering*, 10(5):446–467, 1998.

[12] Chu, W. W., Hsu, C. C. and Taira, R. K. A Unified Timeline Model and User Interface for Multimedia Medical Databases. In *Multimedia Data Management*, pages 149–190, 1998.

[13] Chu, W. W., Hsu, C. C., Cardenas, A. F. and Taira, R. K. Knowledge-Based Image Retrieval with Spatial and Temporal Constructs. *IEEE Transactions on Knowledge and Data Engineering*, 10(6):872–888, November/December 1998.

[14] Chu, W. W., Hsu, C. C., Ieong, I. T. and Taira, R. K. A Temporal Evolutionary Object-Oriented Data Model and Its Query Language for Medical Image Management. In *18th VLDB Conference*, pages 53–64, 1992.

[15] Chu, W. W., Leong, I. and Taira, R. K. A semantic Modeling Approach for Image Retrieval by Content. *VLDB Journal*, 3(4):445–477, 1994.

[16] Chu, W. W., Taira, R. K., Cardenas, A. F., Stepczyk, F. M. and Materna, A. T. Integration and Interoperability of a Multimedia Medical Distributed Database System. *Bulletin of the Technical Commitee on Data Engineering*, 16(1):43–47, 1993.

[17] Comaniciu, D., Meer, P. and Foran, D. J. Image Guided Decision Support System for Pathology. *Machine Vision and Applications*, 11(4):213–224, 1999.

[18] Comaniciu, D., Meer, P. and Foran, D. J. Shape-Based Indexing and Retrieval for Diagnostic Pathology. In *14th International Conference on Pattern Recognition*, pages 902–905, Brisbane, Australia, August, 1998.

[19] Comaniciu, D., Meer, P., Foran, D. J. and Medl, A. Bimodal System for Interactive Indexing and Retrieval of Pathology Images. In *Proc. 4th IEEE Workshop on Applications of Computer Vision (WACV'98)*, pages 76–81, Princeton, NJ, October 1998.

[20] Consorti, F., Merialdo, P. and Sinddoni, G. Metadata Reference Model for Medical Documentation: A Hypermedia Proposal. In *IEEE Metadata Conference*, April 1996.

[21] Dahmen, J., Theiner, T., Keysers, D., Ney, H., Lehmann, T. and Wein, B. Classification of Radiographic in the Image Retrieval in Medical Application System (IRMA). In *Procs 6th International RIAO Conference on Content-Based Multimedia Information Access*, pages 551–566, Paris, France, 2000.

[22] Deng, Y. and Manjunath, B. S. An Efficient Low-Dimensional Color Indexing Scheme for Region-Based Image Retrieval. In *IEEE Intl. Conf on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[23] Dick, K. *XML a Manager's Guide.* Addison Wesley, 2000.

[24] Duda, R. O. and Hart, P. E. *Pattern classification and Scene Analysis.* Wiley, New York, 1973.

[25] Ennis, D. M. and Johnson, N. L. Thurstone-shepard similarity models as special cases of moments generating functions. *Journal of Mathematical Psychology*, 37:104–110, 1993.

[26] Faloutsos, C., Equitz, W., Flickner, M., Niblack, W., Petkovic, D. and Baeder, R. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, 3(3/4):231–262, 1994.

[27] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Steele, D. and Yanker, P. Query by Image and Video Content: The QBIC system. *IEEE Computer*, 28(9):23–32, 1995.

[28] Forsyth, D. A., Malik, J. , Fleck, M. M., Greespan, H., Leung, T., Belongie, S., Carson, C. and Bregler, C. Finding Pictures of Objects in Large Collections of Images. In *Object Representation in Computer Vision*, pages 335–360, 1996.

[29] Furth, B., Smoliar, S. W. and Zhang, H. *Video and Image Processing in Multimedia Systems.* Kluwer Academic Publishers, 1995.

[30] Graham, I. S. *HTML Source Book.* Addison Wesley & Sons, 1997.

[31] Gutman, A. R-tree: A Dynamic Index Structure for Spatial Searching. In *Proc. ACM SIGMOD Conference*, pages 47–57, Boston, MA, June 1984.

[32] Haacke, E. M., Brown, R. W., Thompson, M. R. and Venkatesan, R. *Magnetic Resonance Imaging: Physical Principles and Sequence Design.* Wiley Europe, July 1999.

[33] Hafner, J., Sawhney, H. S., Equitz, W., Flickner, M. and Niblack, W. Efficient Color Histogram Indexing for Quadratic Form Distance Functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7, 1995.

[34] Hagen-Ansert, S. L. *Textbook of Diagnostic Ultrasonography*, volume I. Mosby, 5th edition, January 2001.

[35] Heinrichs, A., Koubaroulis, D., Levienaise-Obadia, B., Rovida, P. and Jolion, J. M . Image Indexing and Content-based Search using Pre-attentive Similarities. In *RIAO 2000*, pages 1616–1631, April 2000.

[36] J. David, N. Dionisio and A. F. Cárdenas. Mquery: A visual query language for multimedia, timeline and simulation date. *Journal of Visual Language and Computing*, 7(4):377–401, 1996.

[37] Kak, A. and Pavlopoulou, C. Computer Vision Techniques for Content-Based Image Retrieval from Large Medical Databases. In *7th Workshop on Machine Vision Applications, IAPR*, Tokyo, Japan, 2000.

[38] Kak, A. C. and Slaney, M. *Principles of Tomographic Imaging.* IEEE Press, 1999.

[39] Kherfi, M. L., Ziou, D. and Bernardi, A. Content-Based Image Retrieval Using Positive and Negative Examples. *To appear.*

[40] Kherfi, M. L., Ziou, D. and Bernardi, A. Web Images Search Engines: A Survey. Technical Report 276, DMI, Faculté des Sciences, Université de sherbrooke, December 2001.

[41] Larsen, J., Hansen, L. K., Szymkowiak, A., Christiansen, T. and Kolenda, T. Webmining: Learning from the World Wide Web. In *Proc. of Nonlinear Methods and Data Mining*, pages 106–125, Rome, Italy, 2000.

[42] Lee, S. Y. and Hsu, F. J. Spatial Reasoning and Similarity Retrieval of Images using 2D C-Sstrings Knowledge Representation. *Pattern Recognition*, 25:305–318, 1992.

[43] Liu, Y. and Dellaert, F. Classification-Driven Medical Image Retrieval. In *Proc. of the Image Understanding Workshop*, 1998.

[44] Liu, Y. and Dellaert, F. Classification Driven Semantic Based Medical Image Indexing and Retrieval. Technical Report CMU-RI-TR-98-25, Robotics Institute, Carnegie Mellon University, 1998.

[45] Liu, Y., Dellaert, F., Rothfus, W. E., Moore, A., Schneider, J. and Kanade, T. Classification-Driven Pathological Neuroimage Retrieval Using Statistical Asymmetry Measure. In *Proc. of the Image Computing and Computer Assisted Intervention Conference (MICCAI)*, Utrecht, The Netherlands, October 2001.

[46] Liu, Y., Rothfus, W. E. and Kanade, T. Classification Driven Semantic Based Medical Image Indexing and Retrieval. Technical Report CMU-RI-TR-98-25, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 1998.

[47] McCamy, C. S., Marcus, H. and Davidson, J. G. A Color-Rendition Chart. *Journal of Applied Photographic Engineering*, 2(3):95–99, 1976.

[48] Muller, W., Muller, H., Marchand-Maillet, S., Pun, T., Squire, D. and Pecenovic, Z. MRML: A communication Protocol for Content-Based Image Retrieval. In *International*

Conference on Visual Information Systems (Visual 2000), Lyon, France, November 2-4, 2000.

[49] Ng, R. and Sedighian, A. Evaluating Multi-dimensional Indexing Structures for Images Transformed by Principal Component Analysis. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, pages 50–61, 1996.

[50] Ogle, V. E. and Stonebraker, M. Chabot: Retrieval from a Relational Database of Images. *IEEE Computer*, 28(9):40–48, 1995.

[51] Orphanoukadis, S. C. and Chronaki, C. Image Retrieval by Pictorial Content in Medical Image Databases. *ERCIM News*, 18:20–21, 1994.

[52] Orphanoukadis, S. C., Chronaki, C. and Kostomanolakis, S. I^2C: A System for the Indexing, Storage and Retrieval of Medical Images by Content. *Medical Informatics*, 12(2):109–122, 1994.

[53] Orphanoukadis, S. C., Chronaki, C. and Zabulis, X. I^2Cnet Medical Image Annotation Service. *Medical Informatics*, 22(4):337–347, 1997.

[54] Orphanoukadis, S. C., Chronaki, C. and Zabulis, X. The I^2Cnet Service Architecture Paradigm. In *MIE'97*, pages 569–600, Porto Carras, Greece, July 2-6, 1995.

[55] Orphanoukadis, S. C., Chronaki, C. and Zabulis, X. Maintaining Medical Image Annotations in I^2Cnet. In *EuroPACS'96*, pages 141–145, Heraklion, Crete, Greece, October 3-5, 1996.

[56] Orphanoukadis, S. C., Kostomanolakis, S. and Chronaki, C. I^2Cnet: Content-Based Similarity Search in Geographically Distributed Repositories of Medical Images. *Computerized Medical Imaging and Graphics*, 20(4):193–207, 1996.

[57] Orphanoukadis, S. C., Tsiknakis, M. and Chronaki, C. Virtual Workspaces in I^2Cnet. *The International Journal of Medical Informatics*, 47(1/2):115–119, 1997.

[58] Orphanoukadis, S. C., Tsiknakis, M., Chronaki, C. and Kostomanolakis, S. The Regional Health Telematic System Of Crete. In *Proceedings of HIealth Telamatics '95*, pages 553–558, Naples, Italy, July 2-6, 1995.

[59] Papathomas, T. V., Conway, T. E., Cox, I. J., Ghosn, J.,Miller, M. L., Minka, T. P., Minka, P. N. and Yianilos, P. N. Psychophysical Studies of the Performance of an Image Database Retrieval system. In *S&T/SPIE Symposium on Electronic Imagimg: Science and Technology, Conference on Human Vision and Electronic Imaging III*, 1998.

[60] Pecenovic, Z., Do, M., Ayer, S. and Vetterli, M. New Methods for Image Retrieval. In *ICPS'89 Congress in Exploring New Tracks in Imaging*, pages 242–246, Antwerp, Belgium, September 1998.

[61] Petrakis, E. and Faloutsos, C. Similarity searching in Medical Image Databases. *IEEE Trans. Knowledge and Data Eng.*, 9(3):435–447, June 1997.

[62] Petrakis, E. and Orphanoudakis, S. Methodology for the Representation, Indexing and Retrieval of Images by Content. *Image and Vision Computing*, 11:504–521, 1993.

[63] Petrakis, E. G. M. and Faloutsos, C. Similarity Searching in Large Image Databases. Technical Report CS-TR-3388, Department of Computer science, University of Maryland, 1994.

[64] Picard, R. W. and Minka, T. P. Vision Texture for Annotation. *Multimedia Systems*, 3(1):3–14, 1995.

[65] Picard, R. W., Minka, T. P. and Szummer, M. Modeling User Subjectivity in Image Libraries. Technical Report 382, MIT Media Laboratory Perceptual Computing Section, September 1996.

[66] Qi, H. and Snyder, W. E. Content-Based Image Retrieval in PACS. *Journal of Digital Imaging*, 12(2):81–82, May 1999.

[67] Rui, Y. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, April 1999.

[68] Rui, Y. and Huang, T. S. Image retrieval: Past, Present, and Future. In *International Symposium on Multimedia Information*, Taipei, Taiwan, December 11-13, 1997.

[69] Rui, Y., Huang, T. S., Ortega, M. and Mehrotra, S. Content-Based Image Retrieval with Relevance Feedback in Mars. In *IEEE International Conference On Image Proc.*, pages 815–818, 1997.

[70] Rui, Y., Huang, T. S., Ortega, M. and Mehrotra, S. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, 1998.

[71] Santini, S. and Jain, R. Similarity Queries in Image Databases. *Lecture Notes in Computer Science*, 1035:571–581, 1996.

[72] Santini, S. and Jain, R. Similarity Measures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(9):871–883, September 1999.

[73] Sclaroff, S. World Wide Web Image Search Engines. Technical Report 1995-016, Image and Video Computing Group, Department of Computer Science, Boston University, 1995.

[74] Sclaroff, S., Cascia, M. L. and Sethi, S. Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. Technical Report 1998-004, Image and Video Computing Group, Department of Computer Science, Boston University, 1998.

[75] Sclaroff, S., Cascia, M. L. and Taycher, L. Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine. Technical Report 1997-014, Image and Video Computing Group, Comp. Sci, Boston University, 1997.

[76] Sclaroff, S., Cascia, M. L. and Taycher, L. ImageRover: A content-Based Image Browser for the World Wide Web. Technical Report 1997-005, Image and Video Computing Group, Department of Computer Science, Boston University, 1997.

[77] Shyu, C., Brodley, C., Kak, A., Broderick, L. S. and Dy, J. G. Testing for Human Perceptual Categories in a Physician-in-the-loop CBIR System for Medical Imagery. In

Proceedings IEEE Workshop of Content-Based Access of Image and Video Databases, pages 25–29, Fort Collins, CO, June 1999.

[78] Shyu, C., Brodley, C., Kak, A., Dy, J., Broderick, L. and Aisen, A. M. Content-Based Retrieval from Medical Image Databases: A Synergy of Human Interaction, Machine Learning and Computer Vision. In *Proc. of the Sixteenth National Conference on Artificial Intelligence*, pages 760–767, Orlando, FL, July 1999.

[79] Shyu, C., Brodley, C., Kak, A., Kosaka, A., Aisen, A. M. and Broderick, L. S. ASSERT: A physician-in-the-loop Content-Based Retrieval System for HRCT Image Databases. *Computer Vision and Image Understanding*, 75(1/2):111–132, July/August 1999.

[80] Shyu, C., Brodley, C., Kak, A., Kosaka, A., Aisen, A. M. and Broderick, L. S. Local versus Global Features for Content-Based Image Retrieval. In *Proceedings IEEE Workshop of Content-Based Access of Image and Video Libraries*, Santa Barbara, CA, June 1998.

[81] Shyu, C., Cai, T. T. and Broderick, L. S. On Archiving and Retrieval of Sequential Images from Tomographic Databases in PACS. In *SPIE Storage and Retrieval for Image and Video Databases VI*, Santa Barbara, CA, January, 1999.

[82] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R. Content-Based Image Retrieval at the End of the Early Years. *Pattern Recognition and Machine Intelligence*, 22(12):1349–1380, December 2000.

[83] Smith, J. R. Image Retrieval Evaluation. In *IEEE Workshop on Content-Based Access of Image and Video Libraries CBAIVL-98*, June 1998.

[84] Smith, J. R. and Chang, S. Automated Image Retrieval Using Color and Texture. Technical Report TR-414-95-20, Columbia University, July 1995.

[85] Smith, J. R. and Chang, S. VisualSEEK: a Fully Automated Content-Based Query System. In *ACM Multimedia*, pages 97–98, 1996.

[86] Smith, J. R. and Chang, S. MetaSEEK: A Content-Based Meta-Search Engine for Images. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 84–95, 1997.

[87] Smith, J. R. and Chang, S. Searching for Images and Videos on the World Wide Web. Technical Report 45996-25, Department of Electrical Engineering and Center for Image Technology for New Media, Columbia University, August 1997.

[88] Squire, D. M., Muller, H., Muller, W. and Pun, T. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. *Pattern Recognition Letters*, 22(5):593–601, 2001.

[89] Squire, D. M., Muller, W., Muller, H. and Raki, J. Content-Based Query of Image Databases, Inspirations from Text Retrieval: Inverted Files, Frequency-Based Weights and Relevance Feedback. In *The 10th Scandinavian Conference on Image Analysis (SCIA'99)*, pages 1–7, June 1999.

[90] Srihari, R. K. Automatic Indexing and Content-Based Retrieval of Captioned Images. *IEEE Computer*, 28(9):49–56, 1995.

[91] Tagare, H. D., Jaffe, C. C. and Ducan, J. Medical Image Databases: A Content-based Retrieval Approach. *Journal of the American Medical Informatics Association*, 4(3), May/Jun 1997.

[92] Thurstone, L. L. A law of comparative judgement. *Psychological Review*, 34:273–286, 1927.

[93] Tversky, A. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[94] Veltkamp, R. C. and Tanase, M. Content-Based Image Retrieval Systems. Technical Report UU-CS-2000-34, Department of Computing Science, Utrecht University, March 2001.

[95] Wang, J., Wang, W. J. and Acharya, R. Color Clustering Techniques for Color Content-Based Image Retrieval from Image Databases. In *Proc. IEEE Conf. on Multimedia Computing and Systems*, 1997.

[96] Wang, J. Z. Pathfinder: Multiresolution Region-Based Searching of Pathology Images using IRM. *Journal of American Medical Informatics Association*, pages 883–887, November 2000.

[97] Wang, J. Z. SIMPLIcity: A Region-Based Image Retrieval System for Picture Libraries and Biomedical Image Databases. In *ACM Multimedia*, pages 483–484, Los Angeles, October 2000.

[98] White, D. and Jain, R. Algorithms and strategies for similarity retrieval. In *TR VCL-96-101*, University of California, San Diego, 1996.

[99] White, D. and Jain, R. Similarity Indexing: Algorithms and performance. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*, 1996.

[100] Wood, M.E.J., Campbell, N. W. and Thomas, B. T. Iterative Refinement by Relevance Feedback in Content-Based Image Retrieval. In *ACM Multimedia 98*, pages 13–20, Bristol, UK, September 1998.

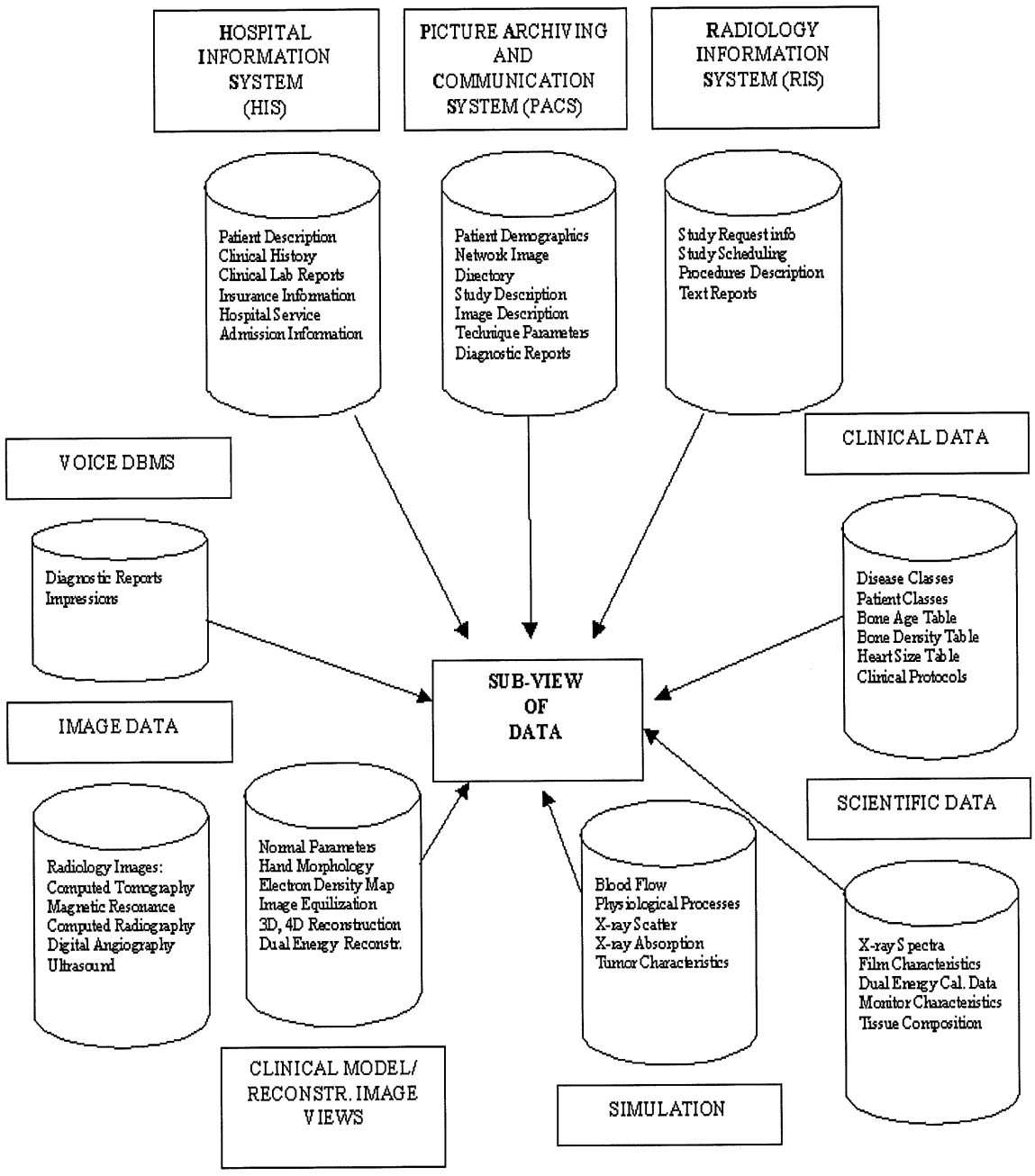


Figure 1: Medical Center multimedia database.

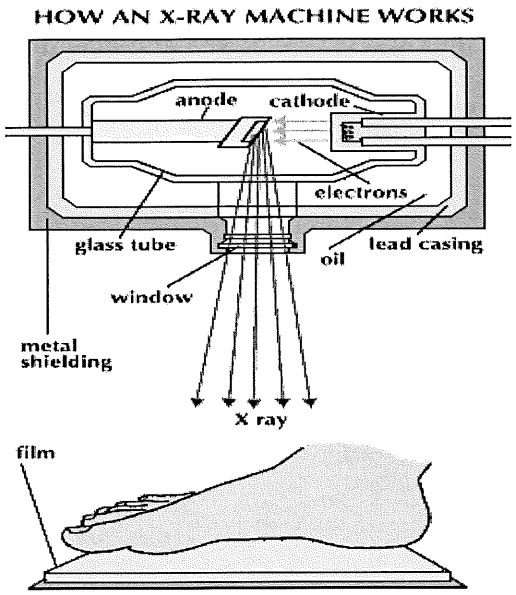


Figure 2: An X-Ray machine sends X-rays from a source within a glass tube through part of the patient's body, behind which is a photographic plate.

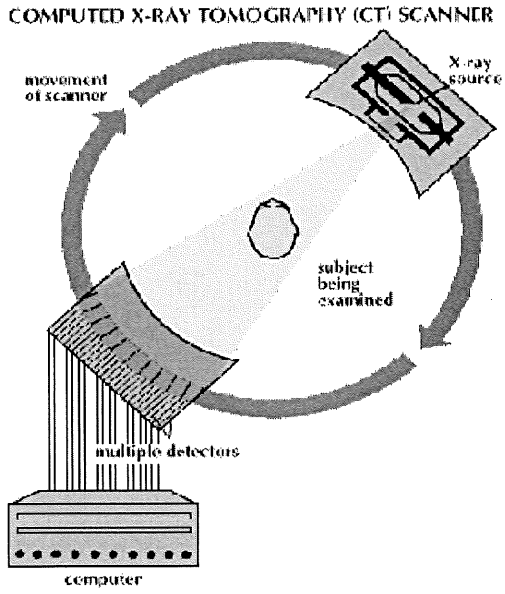


Figure 3: To make a CT scan, a narrow beam of X-rays sweeps across an area of the body, moving through a slight angle after each X-ray pulse. Using the resulting images, a computer produces a three-dimensional X-ray image of the body site.

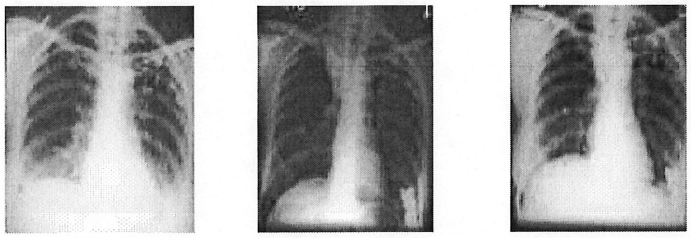


Figure 4: Human chest X-ray images showing dense objects with lighter brightnesses and objects that are less dense with darker brightnesses. The three images show the progressive collapse of the right upper lobe over a period of months.

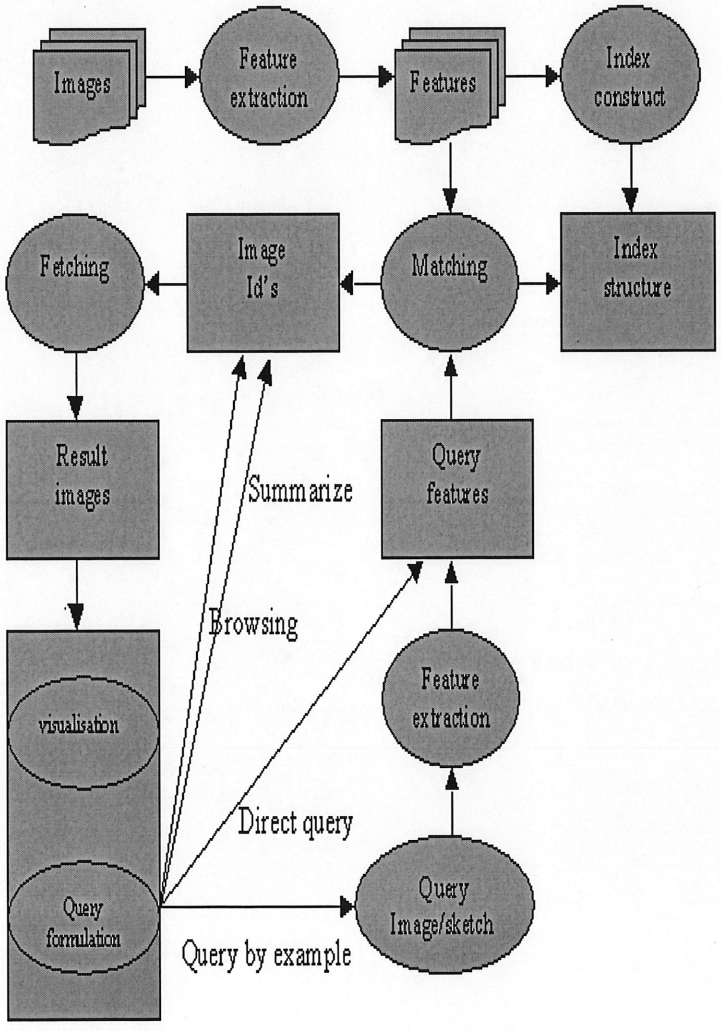


Figure 5: Content-based image retrieval framework.

XML: Easy to Search

<SYMPTOM>fever</SYMPTOM>

<DIAGNOSIS>pharyngitis</DIAGNOSIS>

<Medical.Record.Number>123456789</Medical.Record.Number>

<Phone.Number>210.123.4567</Phone.Number>

Figure 6: XML easy to search

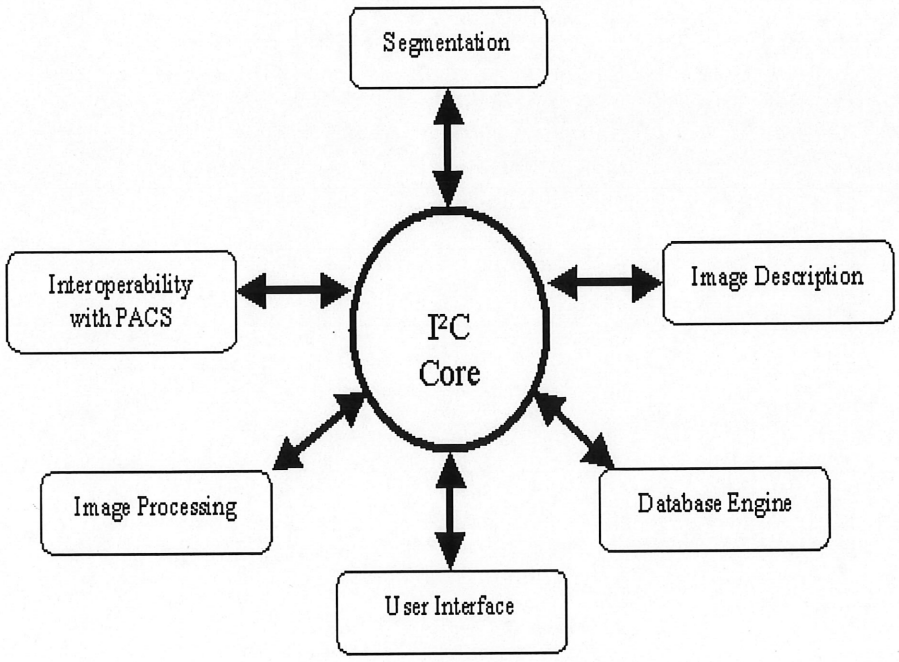


Figure 7: The role of I^2C as an added-value PACS subsystem. The architecture of I^2C is modular. Different modules communicate by exchanging messages through the I^2C core.

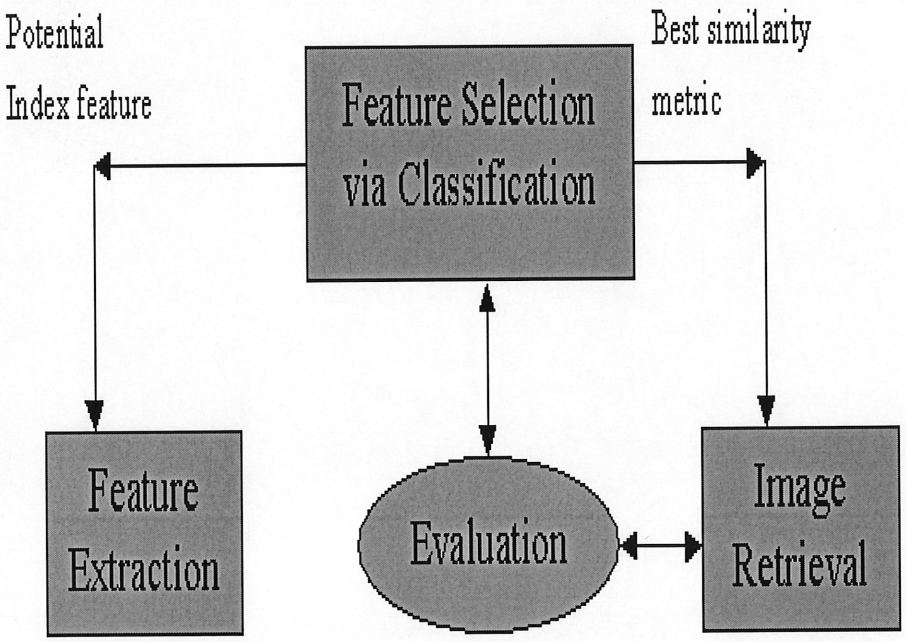


Figure 8: Three steps in classification-driven image retrieval.

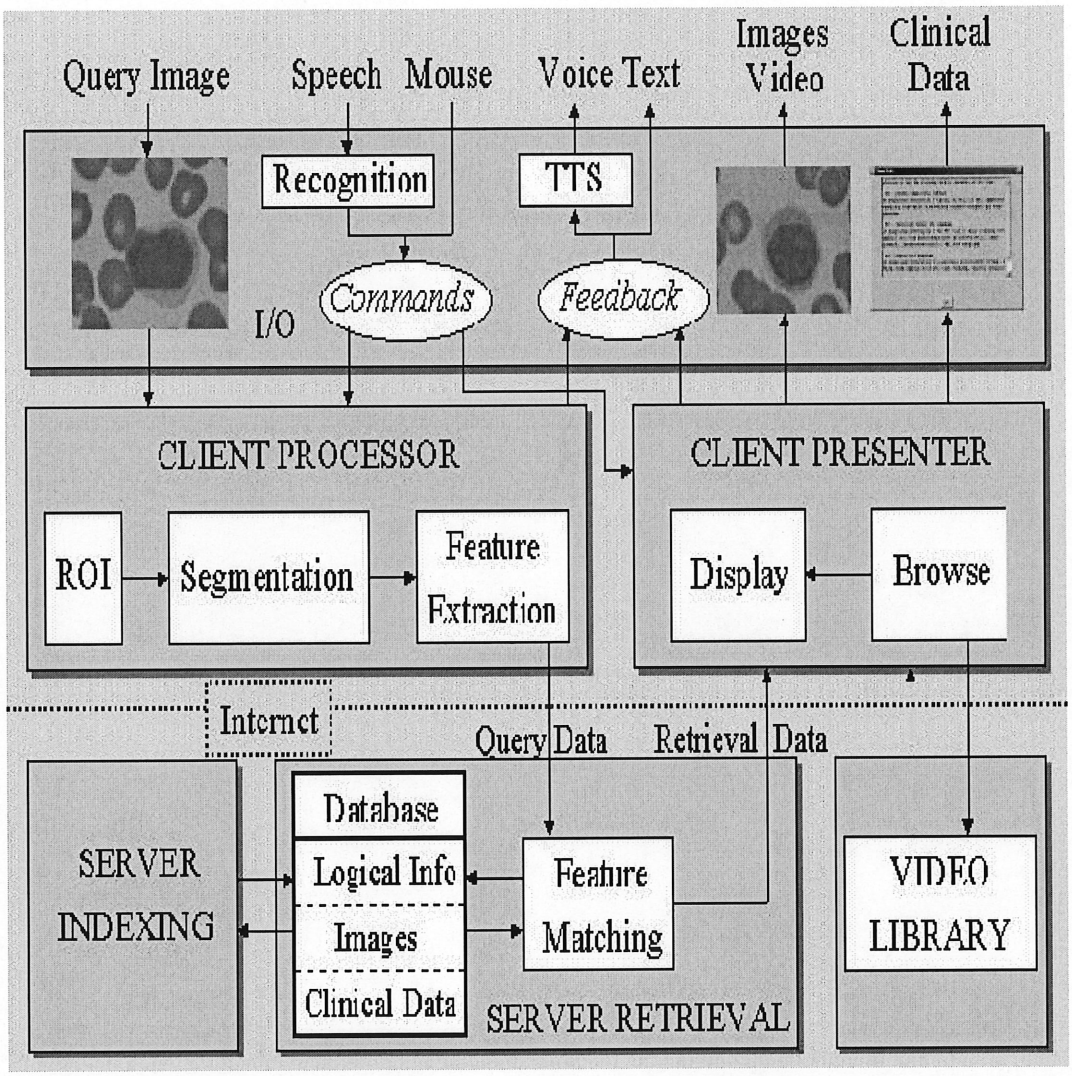


Figure 9: Architecture of the IGDS system.

Les pages 50 à 53 sont manquantes dans notre exemplaire.

Chapitre 2

L'apport des mixtures de Dirichlet dans la classification et la recherche d'images

La deuxième partie de ce mémoire concerne le travail intitulé **Maximum Likelihood Estimation of the Generalized Dirichlet Mixture**. Ce travail concerne une nouvelle mixture, qui peut être appliquée pour la classification et la segmentation des images médicales, utilisant la distribution de Dirichlet. En effet, la majorité des méthodes actuelles utilisant les mixtures ne considèrent que la Gaussienne comme distribution. Toutefois, cette distribution n'est pas le bon choix dans toutes les applications. Ceci est dû au fait que la normale est symétrique et ainsi a une forme rigide. De plus, elle n'a pas un support compact puisqu'elle est définie sur $\mathbb{R}$ ce qui cause l'imprécision des valeurs estimées. Nous pensons que la distribution de Dirichlet est le meilleur choix pour remédier aux inconvénients de la Gaussienne. En effet, la Dirichlet est la généralisation multidimensionnelles de la Beta et elle permet une grande flexibilité. Contrairement à d'autres distributions telle que la normale, la Dirichlet permet différentes formes symétriques et asymétriques. Cette flexibilité a été prouvé par A. El Zaart et D. Ziou [5] qui ont proposé un système nommé GGBL. Ce système est composé de quatre distributions paramétriques (normale, Gamma, Beta et Log-Normal). Ce système est défini par un graphe de chaque distribution dans le plan (β_1, β_2) où β_1 et β_2 sont respectivement le troisième

et le quatrième moment et représentent les coefficients d'asymétrie et de planéité d'une distribution donnée. Dans ce plan, le graphe de la Gaussienne est un point. Le graphe de la Gamma est une ligne. Le graphe de la Beta est un plan. Le graphe du Log-Normal est une ligne. Donc, on peut déduire que la Beta ajuste mieux les données que les autres distributions. Nous avons proposé un algorithme pour l'estimation des paramètres d'une mixture de GDD (une généralisation de la Dirichlet). Cet algorithme est évalué par des méthodes contextuelles et d'autres non-contextuelles. En particulier, nous l'avons utilisé pour le résumé des bases de données d'images.

L'idée de base du problème a été proposée par les Professeurs Djemel Ziou et Jean Vaillancourt, et les recherches nécessaires à la modélisation du problème étaient sous leur direction. Ce travail a fait l'objet d'un rapport [1] qui apparaît dans les pages suivantes de ce mémoire. Une version compacte de cet article [3] est soumise à Computer Vision and Pattern Recognition (CVPR 2003) qui se tiendra à Madison en 2003.

Maximum Likelihood Estimation of the Generalized Dirichlet Mixture [1]

N. Bouguila[(1)], D. Ziou[(1)] and J. Vaillancourt[(2)]

(1) DMI, Faculté des sciences
Université de Sherbrooke
Sherbrooke, Qc, Canada J1K 2R1.

(2) Université du Québec à Hull
283, boulevard Alexandre-Taché
Hull, Qc, Canada J8X 3X7.

Abstract

The Dirichlet distribution offers high flexibility for modeling data. However, it has certain characteristics which present a handicap in practical terms. This paper describes a generalization of the Dirichlet distribution to overcome this handicap which we call the GDD (Generalized Dirichlet Distribution). We propose a method for estimating the parameters of a GDD mixture. This estimation is based on the Maximum Likelihood (ML) and Fisher's scoring methods. A very interesting interpretation, based on the Statistical Geometric Information, is given. To allow convergence to better local minimum and take the GDD distribution into account from the outset, an initialization algorithm is proposed. The performance of our method is tested by contextual and non-contextual evaluation. The non-contextual evaluation is based on synthetic histograms, while the contextual one compares the performance of Gaussian and GDD mixtures in the classification of several pattern-recognition data sets. The GDD mixture is also applied to the problem of summarizing image databases.

Keywords : Dirichlet distribution, mixture modeling, Maximum Likelihood, Fisher's scoring method, Riemannian space, Natural Gradient, image summarizing.

[1]The completion of this research was made possible thanks to Bell Canada's support through its Bell University Laboratories R&D program.

Contents

List of Figures

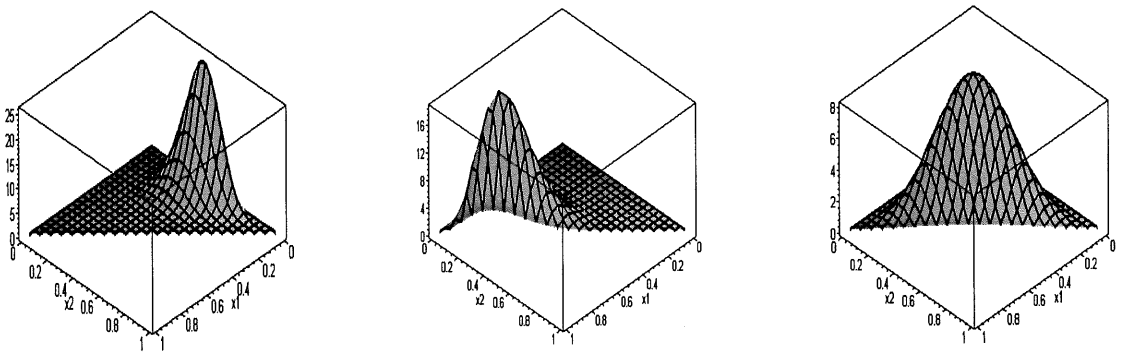
List of Tables

1 Introduction

Scientific pursuits and human activity in general generate data. These data may be incomplete, redundant or erroneous [10]. Probabilistic methods are particularly useful in understanding the patterns present in such data. One such methods is the Bayesian approach which can be roughly described as estimating the uncertainty of a model [36]. In fact, by the Bayesian approach we can estimate the uncertainty of a model's fit and the uncertainty of the estimated parameters themselves. The Bayesian approach can be employed with mixture models, which have been used extensively to model a wide variety of important practical situations where data can be viewed as arising from several populations mixed in varying proportions. Nowadays, this kind of statistical model is used in a variety of domains. In computer vision applications, for example, we can use mixture models to organize image collections as well as for color image segmentation, restoration and texture processing, and content-based image retrieval. Mixture modeling can be viewed as the superimposition of a finite number of component densities. The problem of estimating the parameters of the components of a mixture has been the subject of diverse studies [13]. The isotropic nature of Gaussian functions, along with their capability for representing the distribution compactly by a mean vector and covariance matrix, have made Gaussian Mixture Decomposition (GM) a popular technique [24]. The Gaussian mixture is not the best choice in all applications, however, and it will fail to discover *true* structure where the partitions are clearly non-Gaussian [35, 31]. This is due to the fact that the Gaussian shape is *rigid*, as we will explain below. Moreover, this distribution is defined on $\mathbb{R}$ and thus does not have a compact support, which is why parameters estimated by the moment method, for example, such as the mean, are not accurate in many cases. Indeed, having a compact support is an interesting property for a given density because of the nature of data in general [15]. Generally, we estimate data which are compactly supported, such as data originating from videos, images or text.

In this paper we will show that the Dirichlet distribution can be a very good choice to overcome the disadvantages of the Gaussian. The Dirichlet distribution is the multivariate generalization of the Beta distribution, which offers considerable flexibility and ease of use. In contrast with other distributions such as the Gaussian, which permit only symmetric modes, the Dirichlet

distribution is highly flexible and permit multiple symmetric and asymmetric modes. In fact, the Dirichlet distribution may be skewed to the right, skewed to the left or symmetric (see Fig. 1). The flexibility of the Beta distribution was also proven by A. El Zaart and D. Ziou [12]. In fact, they proposed a system called GGBL, composed of four parametric distributions (Gaussian, Gamma, Beta and Log-Normal). This system was defined by the graph of each distribution in the (β_1, β_2) plane where β_1 and β_2 are respectively the third and the fourth moment and represent the coefficients of the asymmetry and flatness of a given distribution. In this plane, the graph of the Gaussian distribution is represented by a point, that of the Gamma distribution is a line, that of the Beta distribution is a plane and that of the Log-Normal is a line. Thus we can deduce that the Beta distribution can fit data better than the other distributions, particulary the Gaussian and the Gamma.



(a) (b) (c)

Figure 1: The Dirichlet distribution for different parameters. (a) $\alpha_1 = 8.5$, $\alpha_2 = 7.5$, $\alpha_3 = 1.5$. (b) $\alpha_1 = 10.5$, $\alpha_2 = 3.5$, $\alpha_3 = 3.5$. (c) $\alpha_1 = 3.5$, $\alpha_2 = 3.5$, $\alpha_3 = 3.5$.

For all these reasons, we are interested in the Dirichlet distribution. In contrast to the vast amount of theoretical work that exists on the Dirichlet distribution, however, very little work has been done on its practical applications, such as parameter estimation. The majority of the works either consider a single distribution [27, 37] or is restricted to the 2-parameter Beta distribution [14, 16, 4, 39]. This neglect may be due to the fact that this distribution is unfamiliar to many scientists. In this paper, we will propose a generalization of the Dirichlet distribution which we will call the GDD (Generalized Dirichlet Distribution). We will estimate the parameters of a GDD mixture and test it with real data.

The paper is organized as follows. The next section describes the GDD mixture in details. In section 3, we propose a method for estimating the parameters of a GDD mixture. In section 4, we present a way of initializing the parameters and give the complete estimation algorithm. Section 5 is devoted to experimental results. We end the paper with some concluding remarks.

2 The Generalized Dirichlet Distribution Mixture

If the random vector $\vec{X} = (X_1, \ldots, X_{dim})$ follows a Dirichlet distribution [22, 23] the joint density function is given by:

$$p(X_1, \ldots, X_{dim}) = \frac{\Gamma(|\vec{\alpha}|)}{\prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \tag{1}$$

where

$$\sum_{i=1}^{dim} X_i < 1 \tag{2}$$

$$|\vec{X}| = \sum_{i=1}^{dim} X_i, \qquad 0 < X_i < 1 \quad \forall i = 1 \ldots dim \tag{3}$$

$$X_{dim+1} = 1 - |\vec{X}| \tag{4}$$

$$|\vec{\alpha}| = \sum_{i=1}^{dim+1} \alpha_i, \qquad \alpha_i > 0 \quad \forall i = 1 \ldots dim + 1 \tag{5}$$

This distribution is the multivariate extension of the 2-parameter Beta distribution. The mean and the variance of the Dirichlet distribution are given by:

$$E(X_i) = \frac{\alpha_i}{|\vec{\alpha}|} \tag{6}$$

$$Var(X_i) = \frac{\alpha_i(|\vec{\alpha}| - \alpha_i)}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)} \tag{7}$$

and the covariance between X_i and X_j is:

$$Cov(X_i, X_j) = \frac{-\alpha_i \alpha_j}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)} \tag{8}$$

The Dirichlet distribution with parameter vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{dim+1})$ can be represented either as a distribution on the hyperplane $B_{dim+1} = \{(X_1, \ldots, X_{dim+1}), \sum_{i=1}^{dim+1} X_i = 1\}$ in $\mathbb{R}_+^{dim+1}$, or

as a distribution inside the simplex $A_{dim} = \{(X_1, \ldots, X_{dim}), \sum_{i=1}^{dim} X_i < 1\}$ in $\mathbb{R}_+^{dim}$.
This simplex represents a real handicap for us. Indeed, we can't be sure that the data we use will be inside it (between 0 and 1). A simple solution is to apply some transformation, such as normalization, to the vectors $\vec{X} = (X_1, \ldots, X_{dim})$ in order to make them fall inside the simplex A_{dim}. However, much important information may be lost in this type of transformation. Here, we propose a distribution which we call the Generalized Dirichlet Distribution (GDD) in order to overcome this problem. For this purpose, consider $\vec{X} = (X_1, \ldots, X_{dim})$, $0 < X_i < A \ \forall$ $i = 1 \ldots dim$ and $|\vec{X}| < A$. By performing the transformation

$$Y_i = X_i/A, \qquad i = 1, 2, \ldots, dim$$

which has Jacobian A^{-dim}, we obtain: $0 < Y_i < 1 \quad \forall\, i = 1 \ldots dim$ and $|\vec{Y}| < 1$. Suppose that $\vec{Y} = (Y_1, \ldots, Y_{dim})$ follows a Dirichlet distribution with parameter vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{dim+1})$, then:

$$\begin{aligned} p(Y_1, \ldots, Y_{dim}) &= \frac{\Gamma(|\vec{\alpha}|)}{\prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim} Y_i^{\alpha_i - 1} (1 - \sum_{i=1}^{dim} Y_i)^{\alpha_{dim+1} - 1} \\ &= A^{-dim} \frac{\Gamma(|\vec{\alpha}|)}{\prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim} (X_i/A)^{\alpha_i - 1} (1 - \sum_{i=1}^{dim} (X_i/A))^{\alpha_{dim+1} - 1} \\ &= \frac{\Gamma(|\vec{\alpha}|)}{A^{|\vec{\alpha}|-1} \prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim} X_i^{\alpha_i - 1} (A - \sum_{i=1}^{dim} X_i)^{\alpha_{dim+1} - 1} \end{aligned}$$

The GDD is a particular case of the Liouville distribution, where the generating density is uniform and equal to $1/A$ [22] . Thus, if the random vector $\vec{X} = (X_1, \ldots, X_{dim})$ follows a GDD with parameter vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{dim+1})$, the joint density function is given by :

$$p(X_1, \ldots, X_{dim}) = \frac{\Gamma(|\vec{\alpha}|)}{A^{|\vec{\alpha}|-1} \prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \tag{9}$$

This density is defined in the simplex $\{(X_1, \ldots, X_{dim}), \sum_{i=1}^{dim} X_i < A\}$, and we have:

$$X_{dim+1} = A - |\vec{X}| \tag{10}$$

The mean and the variance of the GDD satisfy the following conditions (see Appendix A1):

$$E(X_i) = A \frac{\alpha_i}{|\vec{\alpha}|} \tag{11}$$

$$Var(X_i) = A^2 \frac{\alpha_i(|\vec{\alpha}| - \alpha_i)}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)} \tag{12}$$

and the covariance between X_i and X_j is:

$$Cov(X_i, X_j) = -A^2 \frac{\alpha_i \alpha_j}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)} \tag{13}$$

A GDD mixture with M components is defined as :

$$p(\vec{X}/\Theta) = \sum_{j=1}^{M} p(\vec{X}/j, \Theta_j) P(j) \tag{14}$$

where the $P(j)$ $(0 < P(j) < 1$ and $\sum_{j=1}^{dim} P(j) = 1)$ are the mixing proportions and $p(\vec{X}/j, \Theta_j)$ is the GDD. The symbol Θ refers to the entire set of parameters to be estimated:

$$\Theta = (\vec{\alpha_1}, \ldots, \vec{\alpha_M}, P(1), \ldots, P(M))$$

where $\vec{\alpha_j}$ is the parameter vector for the j^{th} population. In the following developments, we use the notation $\Theta_j = (\vec{\alpha_j}, P(j))$ for j = 1 ... M.

3 Maximum Likelihood Estimation

The problem of estimating the parameters which determine a mixture has been the subject of diverse studies [34]. During the last two decades, the method of maximum likelihood (ML) [7, 38, 33] has become the most common followed approach to this problem. Of the variety of iterative methods which have been suggested as alternatives to optimize the parameters of a mixture, one most widely used is the Expectation Maximization (EM). The EM was originally proposed by Dempster et al. [10] for estimating the Maximum Likelihood Estimator (MLE) of stochastic models. This algorithm gives us an iterative procedure and the practical form is usually very simple. The EM algorithm can be viewed as an approximation of Fisher scoring method [18]. A maximum likelihood estimate associated with a sample of observations is a choice of parameters which maximizes the probability density function of the sample. Thus, with ML estimation, the problem of determining Θ becomes:

$$max_{\Theta} p(\vec{X}/\Theta) \tag{15}$$

with the constraint: $\sum_{j=1}^{M} P(j) = 1$ and $P(j) > 0 \quad \forall j \in [1, M]$ (this constraint is satisfied). These constraints permit us to take into consideration *a priori* probabilities $P(j)$. Using Lagrange multipliers, we maximize the following function:

$$\Phi(\vec{X}, \Theta, \Lambda) = ln(p(\vec{X}/\Theta)) + \Lambda(1 - \sum_{i=1}^{M} P(i)) \tag{16}$$

where Λ is the Lagrange multiplier. For convenience, we have replaced the function $p(\vec{X}/\Theta)$ in Eq. 15 by the function $ln(p(\vec{X}/\Theta))$. If we assume that we have N random vector $\vec{X}_i$ which are independent, we can write:

$$p(\vec{X}/\Theta) = \prod_{i=1}^{N} p(\vec{X}_i/\Theta) \tag{17}$$

$$p(\vec{X}_i/\Theta) = \sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j) P(j) \tag{18}$$

Replacing equations 17 and 18 we obtain:

$$\Phi(\vec{X}, \Theta, \Lambda) = \sum_{i=1}^{N} ln(\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j) P(j)) + \Lambda(1 - \sum_{j=1}^{M} P(j)) \tag{19}$$

We will now try to resolve this optimization problem. To do this, we must determine the solution to the following equations:

$$\frac{\partial}{\partial \Theta} \Phi(\vec{X}, \Theta, \Lambda) = 0 \tag{20}$$

$$\frac{\partial}{\partial \Lambda} \Phi(\vec{X}, \Theta, \Lambda) = 0 \tag{21}$$

Calculating the derivative with respect to Θ_j, we obtain (see Appendix A2):

$$\frac{\partial}{\partial \Theta_j} \Phi(\vec{X}, \Theta, \Lambda) = \sum_{i=1}^{N} p(j/\vec{X}_i, \Theta_j) \frac{\partial}{\partial \Theta_j} ln(p(\vec{X}_i/j, \Theta_j)) \tag{22}$$

where $p(j/\vec{X}_i, \Theta_j)$ is the posterior probability. In what follows, we will estimate the parameters.

3.1 Estimation of the *a priori* probability

Since $p(\vec{X}_i/j, \vec{\alpha_j})$ is independent of $P(j)$, straight forward manipulations yield:

$$\frac{\partial}{\partial P(j)} \Phi(\vec{X}, \Theta, \Lambda) = \frac{1}{P(j)} \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) - \Lambda = 0 \tag{23}$$

Now we take the derivative of Eq. 19 with respect to Λ. We find:

$$\frac{\partial}{\partial \Lambda}\Phi(\vec{X},\Theta,\Lambda) = 1 - \sum_{j=1}^{M} P(j) = 0 \tag{24}$$

From Eq. 23, we obtain:

$$P(j) = \frac{1}{\Lambda}\sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j}) \tag{25}$$

Thus Eq. 24 gives us:

$$\sum_{j=1}^{M} P(j) = \frac{1}{\Lambda}\sum_{j=1}^{M}\sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j}) = 1 \tag{26}$$

Since

$$\sum_{j=1}^{M} p(j/\vec{X}_i,\vec{\alpha_j}) = 1 \tag{27}$$

This gives

$$\sum_{j=1}^{M} P(j) = \frac{1}{\Lambda}\sum_{i=1}^{N}\sum_{j=1}^{M} p(j/\vec{X}_i,\vec{\alpha_j}) = \frac{N}{\Lambda} = 1 \tag{28}$$

Thus,

$$\Lambda = N \tag{29}$$

Finally, the *a priori* probability is:

$$P(j) = \frac{1}{N}\sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j}) \tag{30}$$

3.2 Estimation of the $\vec{\alpha}$ parameters

In order to estimate the $\vec{\alpha}$ parameters we will use Fisher's scoring method. This approach is a variant of the Newton-Raphson [32] method. In fact, Eq. 20 can be approximated by expanding it in a power series around a point Θ_{j_0}:

$$\frac{\partial}{\partial \Theta_j}\Phi(\vec{X},\Theta,\Lambda) \simeq \frac{\partial}{\partial \Theta_j}\Phi(\vec{X},\Theta,\Lambda)(\Theta_{j_0}) + (\Theta_j - \Theta_{j_0})\frac{\partial^2}{\partial \Theta_j^2}\Phi(\vec{X},\Theta,\Lambda)(\Theta_{j_0}) \tag{31}$$

Since $\frac{\partial}{\partial \Theta_j}\Phi(\vec{X},\Theta,\Lambda) = 0$, then

$$\Theta_j \simeq \Theta_{j_0} - (\frac{\partial^2}{\partial \Theta_j^2}\Phi(\vec{X},\Theta,\Lambda)(\Theta_{j_0}))^{-1}\frac{\partial}{\partial \Theta_j}\Phi(\vec{X},\Theta,\Lambda)(\Theta_{j_0}) \tag{32}$$

Thus, an updated estimate, $\hat{\Theta}_j^{(k+1)}$, of a current estimate $\hat{\Theta}_j^{(k)}$ is given by:

$$\hat{\Theta}_j^{(k+1)} = \hat{\Theta}_j^{(k)} - (\frac{\partial^2}{\partial\Theta_j^2}\Phi(\vec{X},\Theta,\Lambda)(\Theta_j))^{-1}_{\Theta_j=\hat{\Theta}_j^{(k)}}(\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda)(\Theta_j))_{\Theta_j=\hat{\Theta}_j^{(k)}} \quad (33)$$

which is equivalent to:

$$\hat{\Theta}_j^{(k+1)} = \hat{\Theta}_j^{(k)} - H^{-1}(\hat{\Theta}_j^{(k)})(\frac{\partial}{\partial\Theta_j}\Phi(\vec{X},\Theta,\Lambda)(\Theta_j))_{\Theta_j=\hat{\Theta}_j^{(k)}} \quad (34)$$

where H is the Hessian matrix evaluated at the current estimate. One variant of this approach is Fisher's scoring method, where the Hessian matrix H is replaced by the negative of Fisher's information matrix. The scoring method is based on the first, second and mixed derivatives of the log-likelihood function. Thus, we will compute these derivatives.

According to Eq. 22 we have:

$$\frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda) = \sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j})\frac{\partial}{\partial\alpha_{jl}}ln(p(\vec{X}_i/j,\vec{\alpha_j})) \quad (35)$$

Calculating the derivative of $ln(p(\vec{X}_i/j,\vec{\alpha_j})$ with respect to Θ_j, we obtain (see Appendix A3):

$$\frac{\partial}{\partial\alpha_{jl}}ln(p(\vec{X}_i/j,\vec{\alpha_j})) = \Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A) \quad (36)$$

where $\Psi(.)$ is the digamma function. Thus:

$$\frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda) = (\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}))\sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j}) + \sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j})(ln(X_{il}) - ln(A)) \quad (37)$$

During iterations, the α_{jl} can become negative. In order to overcome this problem, a suggestion was given by Ronning [37] for the case of one Dirichlet distribution. His suggestion is to set all $\alpha_{jl} = min\{X_{il}\}$, $i = 1 \ldots N$. These initial estimates only prevent the α_{jl} from becoming negative during the first few iterations. Besides, the method gives good results only in the case of one distribution, because of sensitivity to initialization in the case of a mixture (see next section). Here, we give a better solution for keeping the α_{jl} positive during all the iterations. Since we require that the α_{jl} be strictly positive, and we want the parameters upon which we will derive to be unconstrained, we reparametrize, setting $\alpha_{jl} = e^{\beta_{jl}}$, where β_{jl} is an unconstrained real number. Then, the partial derivative of Φ (Eq. 19) with respect to β_{jl} is as follows (see Appendix A4):

$$\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda) = \alpha_{jl}[(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}))\sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j}) + \sum_{i=1}^{N} p(j/\vec{X}_i,\vec{\alpha_j})(ln(X_{il}) - ln(A))] \quad (38)$$

By computing the second and mixed derivatives of the log-likelihood function we obtain (see Appendix A5):

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{jl}^2}\Phi(\vec{X},\Theta,\Lambda) &= \frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda)+\alpha_{jl}^2(\Psi'(|\vec{\alpha_j}|)-\Psi'(\alpha_{jl}))\sum_{i=1}^{N}p(j/\vec{X_i},\vec{\alpha_j})\\ &+ \alpha_{jl}[(\Psi(|\vec{\alpha_j}|)-\Psi(\alpha_{jl}))\sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl}}p(j/\vec{X_i},\vec{\alpha_j})\\ &+ \sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl}}p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il})-ln(A))]\end{aligned} \tag{39}$$

$$\begin{aligned}\frac{\partial^2}{\partial\beta_{jl_1}\partial\beta_{jl_2}}\Phi(\vec{X},\Theta,\Lambda) &= \alpha_{jl_1}\alpha_{jl_2}\Psi'(|\vec{\alpha_j}|)\sum_{i=1}^{N}p(j/\vec{X_i},\vec{\alpha_j})\\ &+ \alpha_{jl_1}[(\Psi(|\vec{\alpha_j}|)-\Psi(\alpha_{jl_1}))\sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl_2}}p(j/\vec{X_i},\vec{\alpha_j})\\ &+ \sum_{i=1}^{N}\frac{\partial}{\partial\beta_{jl_2}}p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il_1})-ln(A))]\end{aligned} \tag{40}$$

where $\Psi'(.)$ is the trigamma function. Note that we need to compute the derivative of the *a posterior* probability $p(j/\vec{X_i},\vec{\alpha_j})$ with respect to β_{jl} (see Appendix A6):

$$\frac{\partial}{\partial\beta_{jl}}p(j/\vec{X_i},\vec{\alpha_j}) = \alpha_{jl}\times p(j/\vec{X_i},\vec{\alpha_j})(1-p(j/\vec{X_i},\vec{\alpha_j}))(\Psi(|\vec{\alpha_j}|)-\Psi(\alpha_{jl})+ln(X_{il})-ln(A)) \tag{41}$$

Given a set of initial estimates (see the next section), Fisher's scoring method can now be used. The iterative scheme of the Fisher method is given by the following equation:

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_{j1}\\ \vdots\\ \hat{\beta}_{jdim+1}\end{pmatrix}^{new} &= \begin{pmatrix}\hat{\beta}_{j1}\\ \vdots\\ \hat{\beta}_{jdim+1}\end{pmatrix}^{old}\\ &+ \begin{pmatrix}Var(\hat{\beta}_{j1}) & \dots & Cov(\hat{\beta}_{j1},\hat{\beta}_{jdim+1})\\ \vdots & \ddots & \vdots\\ Cov(\hat{\beta}_{jdim+1},\hat{\beta}_{j1}) & \dots & Var(\hat{\beta}_{jdim+1})\end{pmatrix}^{old}\begin{pmatrix}\frac{\partial}{\partial\hat{\beta}_{j1}}\Phi\\ \vdots\\ \frac{\partial}{\partial\hat{\beta}_{jdim+1}}\Phi\end{pmatrix}^{old}\end{aligned} \tag{42}$$

where j is the class number: $1\leq j\leq M$.

The variance-covariance matrix is obtained as the inverse[2] of the Fisher's information matrix **I**. The information matrix **I** is:

$$\mathbf{I} = I_{l_1l_2} = -E[\frac{\partial^2}{\partial\beta_{jl_1}\partial\beta_{jl_2}}\Phi(\vec{X},\Theta,\Lambda)] \tag{43}$$

Comparing this iterative scheme based on Fisher's scoring method with a quasi-Newton method presented by the following equation [34, 11]:

$$\beta_{jl}^{new} := \beta_{jl}^{old} - \eta\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda) \tag{44}$$

[2]We have made some approximations to avoid inverting the information matrix in each iteration. See Appendix B for more details.

where $0 < \eta \leq 1$, we can note the presence of an extra term which is the inverse of the Fisher's information matrix. Let us now focus on the geometrical interpretation of Eq. 42:

$$\beta_{jl}^{new} := \beta_{jl}^{old} - \underbrace{\left(\begin{array}{ccc} Var(\hat{\beta}_{jl}) & \dots & Cov(\hat{\beta}_{jl}, \hat{\beta}_{jdim+1}) \end{array} \right)^{old} \left(\begin{array}{c} \frac{\partial}{\partial \hat{\beta}_j} \Phi \\ \vdots \\ \frac{\partial}{\partial \hat{\beta}_{jdim+1}} \Phi \end{array} \right)^{old}}_{NG} \tag{45}$$

Thus, the ordinary gradient $\frac{\partial}{\partial \beta_{jl}} \Phi(\vec{X}, \Theta, \Lambda)$ is replaced by the term NG, which is called the *Natural Gradient* or *Contravariant Gradient* by Amari [3]. Let $S = \{\vec{\beta} = (\beta_1, \dots, \beta_{dim+1}) \in \mathbb{R}^{dim+1}\}$ be a parameter space in which a given likelihood function (Φ) is defined. When S is a Euclidean space with an orthonormal coordinate system, the squared length of a small incremental vector $\partial\vec{\beta} = (\partial\beta_1, \dots, \partial\beta_{dim+1})$ connecting $\vec{\beta}$ and $\vec{\beta} + \partial\vec{\beta}$ is given by:

$$|\partial\vec{\beta}|^2 = \sum_{l_1=1}^{dim+1} (\partial\beta_i)^2 \tag{46}$$

However, when the coordinate system is nonorthonormal, the squared length is given by the following quadratic form [3]:

$$|\partial\vec{\beta}|^2 = \sum_{l_1=1}^{dim+1} \sum_{l_2=1}^{dim+1} g_{l_1 l_2} (\partial\beta_{l_1})(\partial\beta_{l_2}) \tag{47}$$

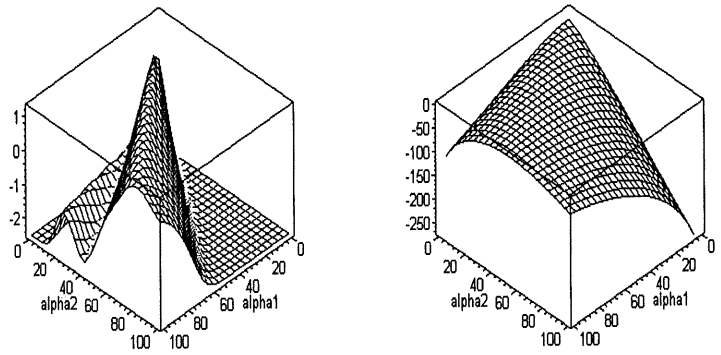
The $(dim + 1) \times (dim + 1)$ matrix $G = (g_{l_1 l_2})$ is called the Riemannian metric tensor. This matrix is reduced to the unit matrix $I_{dim+1 \times dim+1}$ in the Euclidean orthonormal case. Knowing that the GDD is an exponential density [22], we can affirm that the parameter space of our likelihood function Φ, described by Eq. 19, is a curved manifold. In fact, according to Amari, the exponential family of probability forms a manifold which is equipped with a Riemannian metric given by the Fisher's information matrix [2, 3, 1]. Thus, we do not have an orthonormal linear coordinate system, and the length of $\partial\vec{\beta}$ is writen as Eq. 47. Knowing that the steepest direction of a function Φ at $\vec{\beta}$ is defined by the vector $\partial\vec{\beta}$ that minimizes $\Phi(\beta + \partial\vec{\beta})$, where $|\partial\vec{\beta}|$ has a fixed length and is sufficiently small, we can deduce that the gradient of Φ can't be the same in a Euclidean space and a Riemannian one. The relation between the natural gradient $\frac{\partial\check{\phi}}{\partial\beta_l}$ and the ordinary gradient is given by the following equation [2]:

$$\frac{\partial\check{\phi}}{\partial\beta_l} = G^{-1} \frac{\partial\phi}{\partial\beta_l} \tag{48}$$

where G is the Fisher's information matrix. This result was confirmed by experiments. Indeed, we have implemented these two methods and observed that the method given by Eq. 44 does not give good results compared with the Fisher's scoring method.

4 Initialization and Convergence Test

The maximum likelihood function presented by Eq. 19 is globally concave [37] in the case of one distribution (M = 1). However, this particular advantage is not preserved when $M > 1$, as shown in Fig. 2. In order to make our algorithm less sensitive to local maxima, we have used some initialization schemes including the Fuzzy C-means [6] and the method of moments (MM). In fact, the method of moments gives really good estimations because of the compact support of the Dirichlet distribution.



(a) (b)

Figure 2: Representation of Φ as function of two parameters α_1 and α_2. (a) $M > 1$. (b) $M = 1$.

From an examination of Eq. 11 and Eq. 12 we see that there are first *dim* first-order moments and *dim* second-order moments, yielding a total of $C_{dim+1}^{2(dim+1)}$ possible combinations of equations to solve for the *dim* parameters. According to Fieltiz and Myers [14] a symmetrical way of proceeding would be to choose the first *dim* first-order equations and the first second-order equation. The reason for not choosing the $(dim + 1)$-th first order equation is that the $(dim + 1)$-th equation is a linear combination of the others and together they do not form an

independent set of equations. Thus we have:

$$\alpha_l = \frac{(x'_{11} - x'_{21})x'_{1l}}{x'_{21} - (x'_{11})^2} \quad l = 1, 2 \ldots, dim \tag{49}$$

and

$$\alpha_{dim+1} = \frac{(x'_{11} - x'_{21})(1 - \sum_{l=1}^{dim} x'_{1l})}{x'_{21} - (x'_{11})^2} \tag{50}$$

where

$$x'_{1l} = \frac{1}{A \times N} \sum_{i=1}^{N} x_{il} \quad l = 1, 2 \ldots, dim + 1 \tag{51}$$

$$x'_{21} = \frac{1}{A^2 \times N} \sum_{i=1}^{N} x_{i1}^2 \tag{52}$$

Thus, our initialization method can be resumed as follows (we suppose that the number of clusters M is known):

INITIALIZATION Algorithm

1. Apply the Fuzzy C-means to obtain the elements, covariance matrix and mean of each component.
2. Apply the MM for each component j to obtain the vector of parameters $\vec{\alpha_j}$.
3. Assign the data to clusters, assuming that the current model is correct.
4. If the current model and the new model are sufficiently close to each other, terminate, else go to 2.

We can readily note that this initialization algorithm take the distribution into account. In contrast to the *classic* initialization methods which use only algorithms such as K-means to obtain the initialization parameters, we have introduced the method of moments with an iterative scheme to refine the results. By using the method of moments, we suppose from the outset that we have a GDD mixture. This initialization method is designed to work on large databases. When working on small data sets, applying the Fuzzy C-means and the MM only once is a feasible option. With this initialization method in hand, our algorithm for estimating of GDD mixtures can be summarized as follows:

GDD MIXTURE ESTIMATION Algorithm

1. INPUT: dim-dimensional data X_i, $i = 1, \ldots, N$ and the number of clusters M.
2. INITIALIZATION Algorithm.
3. Update the $\vec{\alpha_j}$ using Eq. 42, $j = 1, \ldots, M$.
4. Update the $P(j)$ using Eq. 30, $j = 1, \ldots, M$.
5. If the convergence test is passed, terminate, else go to 3.

If the sample is sufficiently large, the test of convergence can be done using a statistical method [8]. The method uses a quadratic form of the gradient vector. Consider the statistics:

$$S = \left(\frac{\partial}{\partial \hat{\beta}_{j1}}\Phi \quad \cdots \quad \frac{\partial}{\partial \hat{\beta}_{jdim+1}}\Phi \right) \begin{pmatrix} Var(\hat{\beta}_{j1}) & \cdots & Cov(\hat{\beta}_{j1}, \hat{\beta}_{jdim+1}) \\ \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_{jdim+1}, \hat{\beta}_{j1}) & \cdots & Var(\hat{\beta}_{jdim+1}) \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial \hat{\beta}_{j1}}\Phi \\ \vdots \\ \frac{\partial}{\partial \hat{\beta}_{jdim+1}}\Phi \end{pmatrix} \tag{53}$$

This statistical test can be shown to be approximately distributed as a Chi-square random variable with $dim + 1$ degrees of freedom. The iteration is continued until S falls below than $\chi^2_{dim+1}(\nu)$ for a fixed ν. Other convergence tests could involve testing the stabilization of the $\vec{\beta_j}$ or the value of the maximum likelihood function.

5 Experimental Results

In this section, we validate the GDD mixture using contextual and non-contextual evaluation [29] to test the performance of our method. For the non-contextual evaluation, we use some synthetic histograms. The contextual evaluation is based on a pattern recognition application and one from computer vision. We begin with the non-contextual evaluation. For this purpose we synthesized three histograms (figures 3, 4 and 5). The real and the estimated parameters of each histogram are specified in tables 1, 2 and 3. We also defined an error measure, given by:

$$E = \frac{1}{N} \sum_{i=1}^{N} |H_{real}(\vec{X_i}) - H_{estimated}(\vec{X_i})| \tag{54}$$

where

$$H_{real}(\vec{X}) = H_{estimated}(\vec{X}) = \sum_{j=1}^{M} P(j) \frac{\Gamma(|\vec{\alpha}|)}{A^{|\vec{\alpha}|-1} \prod_{i=1}^{dim+1} \Gamma(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \tag{55}$$

N is the number of data point used to compute the error, $H_{real}(\vec{X})$ is the real value of the histogram in $\vec{X}$ and $H_{estimated}(\vec{X})$ is the estimated value. The first histogram presents a GDD mixture of three well separated components. We used $N = 100$, the error was 3.433×10^{-8}. The second and the third histograms present overlapped GDD components. The errors were 5.99×10^{-7} and 2.5×10^{-2}, respectively for $N = 255$.

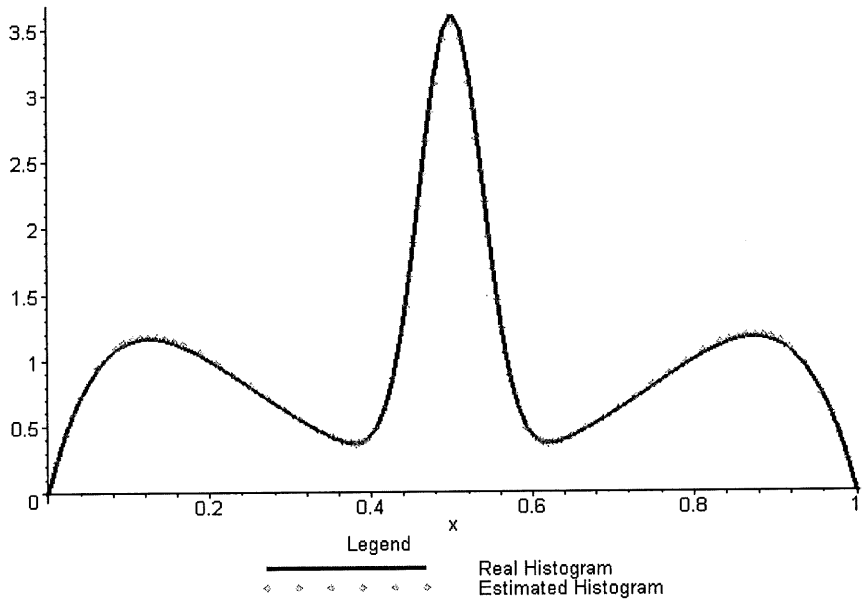


Figure 3: Real and estimated histograms for the first synthetic data set.

In the pattern recognition application, our method was used to model the class-conditional densities in four standard pattern recognition data sets which differ in dimension, size and complexity. The classification was performed using the Bayes rule (X_i is assigned to class j_1 if $P(j_1)p(x_i/j_1) > P(j)p(x_i/j), \forall j \neq j_1$) after the class-conditional densities have been estimated. The goal of this application is also to compare the modeling capabilities of GDD and Gaussian mixture. We have used the EM algorithm to estimate the parameters of Gaussian mixtures because it's very hard in practice to use the scoring method in the case of Gaussians [40, 30]. The comparison will be based essentially on errors of classification, error of fit and number

	Real parameters	Estimated parameters
Mode 1	$P(1)$=0.33	$P(1)$=0.333
	$\alpha_{11} = 8$	$\alpha_{11} = 8.116$
	$\alpha_{12} = 2$	$\alpha_{12} = 2.024$
Mode 2	$P(2)$=0.34	$P(2)$=0.335
	$\alpha_{21} = 80$	$\alpha_{21} = 79.567$
	$\alpha_{22} = 80$	α_{22} =79.567
Mode 3	$P(3)$=0.33	$P(3)$=0.332
	$\alpha_{31} = 2$	$\alpha_{31} = 2.024$
	$\alpha_{32} = 8$	$\alpha_{32} = 8.116$

Table 1: Estimation of the parameters of the GDD for the first synthetic data set.

of iterations in each case. We begin with two examples which are reported in [5, 9]. The first data set describes an enzymatic activity distribution in the blood and the second one an acidity index distribution for 155 lakes. For these two data sets, a mixture of 2 distributions is identified [5, 9]. Figures 6 and 8 show the real and the estimated histograms for the Enzyme and Acidity data sets, respectively. In both cases, it's clear that the GDD and the Gaussian fit the data. We also compared the likelihood cycle of the Gaussian and the GDD (figures 7 and 9). According to these figures our algorithm converges in a smaller number of iterations (9 for the Enzyme data set and 14 for the Acidity set) compared to the case where Gaussian mixture was considered (17 for the Enzyme data set and 20 for the Acidity set). The final results of the estimations are given in tables 4 and 5. Our algorithm was also validated with multidimensional data sets. We took two well-known examples, the Ruspini [19] and Wisconsin Breast Cancer [21] data sets. We chose these data sets for their specific characteristics. The Ruspini data set contains two-dimensional[3] data in four groups (see Fig. 10) and the Breast Cancer data set is characterized by its size (683 patterns) and its dimension (9) [26]. For both

[3]There is of course, no requirement that the problem be of such low dimensionality and we chose $dim = 2$ purely for ease of presentation. It does, however, become increasingly difficult to verify the results of any modeling when the dimensionality is high.

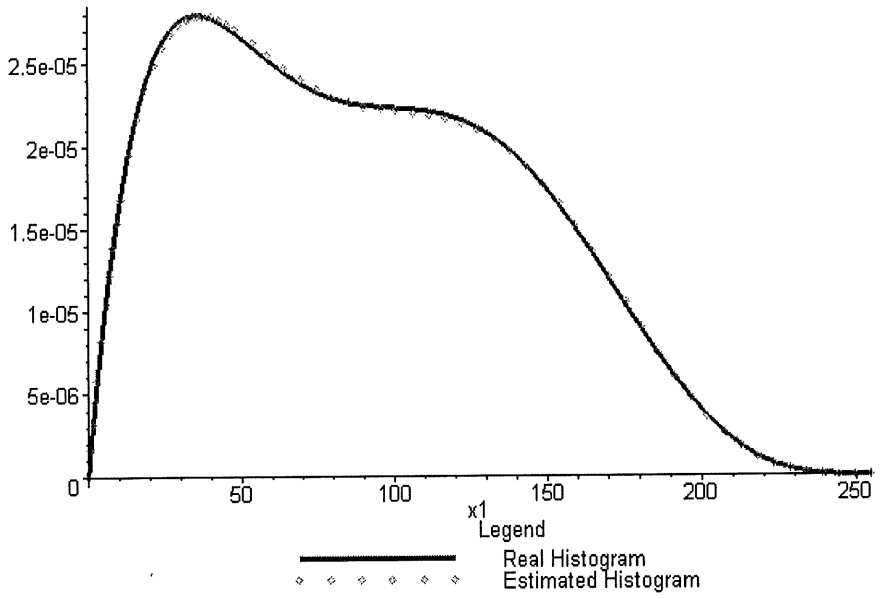


Figure 4: Real and estimated histograms for the second synthetic data set.

examples, the comparison between the GDD and the Gaussian mixtures is based on the errors of classification. By using the GDD mixture for the Ruspini data, we reached convergence in 10 iterations with an error of 1.33 percent (see tables 6 and 7). This is slightly better than the result found for the Gaussian mixture (an error of 2.66 percent in 11 iterations). We also plotted the results. In each case we can clearly observe the presence of 4 classes (see figures 11 and 12). The GDD also gave better results (an error of 1.024 percent) for the Breast Cancer data, compared with the Gaussian mixture (an error of 2.342 percent). The estimated parameters are given in tables 8 and 9.

The third validation is contextual and concerns the summarization of image databases. For this validation, we have used only GDD mixture because of the difficulty to applicate Gaussian mixture (due to the singularity of the covariance matrix during iterations). This application is very important especially in the case of content-based image retrieval [25]. Summarizing the database simplifies the task of retrieval by restricting the search for similar images to a smaller domain of the database. Summarization is also very efficient for browsing [28]. Knowing the categories of images in a given database allows the user to find the images he is looking for more quickly. Using mixture decomposition, we can find natural groupings of images and represent each group by the most representative image in the group. In other words, after appropriate

	Real parameters	Estimated parameters
Mode 1	$P(1)$=0.5	$P(1)$=0.545
	$\alpha_{11} = 2$	$\alpha_{11} = 1.942$
	$\alpha_{12} = 8$	$\alpha_{12} = 7.141$
Mode 2	$P(2)$=0.5	$P(2)$=0.455
	$\alpha_{21} = 5$	$\alpha_{21} = 5.466$
	$\alpha_{22} = 5$	α_{22} =5.108

Table 2: Estimation of the parameters of the GDD for the second synthetic data set.

features are extracted from the images, it allows us to partition the feature space into regions that are relatively homogeneous, with respect to the chosen set of features. By identifying the homogeneous regions in the feature space, the task of summarization is accomplished. We used a database containing 600 images of size 128 $\times$ 96, and took color as a feature for categorizing the images. In order to determine the vector of characteristics for each image, pixels were projected onto the $3D$ HSI (H = Hue, S = Saturation, and I = Intensity) space. We thus obtained a $3D$ color histogram for each image. Based on the work of Kherfi et al. [20], we obtained an $8D$ vector from this histogram. Their method consists of partitioning the space by subdividing each of the axes H, S and I into n equal intervals. This gives n^3 subspaces. The sum of the elements in each subspace is computed and the result is placed in the corresponding cell of the feature vector. In our application, we chose $n = 2$, so each image was represented by a $2^3 = 8D$ feature vector. We also asked a human subject to determine the number of groups, and he found five categories. After the feature were extracted from the images, the GDD mixture algorithm was applied to the feature vectors by specifying five classes, where each vector represents an image. The two classifications (the one generated by the human subject and the one given by our algorithm) were compared by counting the number of misclassified images, yielding the confusion matrix (see table 10). In this confusion matrix, the cell $(classi, classj)$ represents the number of images from $classi$ which are classified as $classj$. Our algorithm reached the convergence in 15 iterations (see Fig. 14). The number of images misclassified was small: 40 images, which represents an accuracy of 93.34 percent.

6 Conclusion

In this paper, we have introduced a new mixture, based on a generalization of the Dirichlet distribution, that we call the GDD. The GDD has the advantage that by varying its parameters, it permits multiple modes and asymmetry and can thus approximate a wide variety of shapes. We estimated the parameters of this mixture using the maximum likelihood and Fisher's scoring methods. An interesting interpretation, based on the statistical geometric information, was given. In order to make our method less sensitive to initialization, we proposed an initialization algorithm based on the moments, which takes the distribution into account from the outset. Contextual and non-contextual evaluations were used to test the performance of our method. The non-contextual evaluation was based on synthetic histograms. The contextual test involved data classification and summarization of image databases. From the results of these evaluations, we can say that the GDD mixture has good modeling capabilities.

A Appendix A

A.1 Appendix A1

Here, we determine the mean, variance and covariance of a GDD. Take the random vector $\vec{X} = (X_1, \ldots, X_{dim})$, $0 < X_i < A \; \forall \; \mathrm{i} = 1 \ldots dim$ and $|\vec{X}| < A$. Suppose that this vector follows a GDD with parameter vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{dim+1})$, then:

$$E(X_i) = E(A \times Y_i) = A \times E(Y_i) = A\frac{\alpha_i}{|\vec{\alpha}|}$$

$$Var(X_i) = Var(A \times Y_i) = A^2 \times Var(Y_i) = A^2\frac{\alpha_i(|\vec{\alpha}| - \alpha_i)}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)}$$

$$Cov(X_i, X_j) = Cov(A \times Y_i, A \times Y_j) = A^2 \times Cov(Y_i, Y_j) = -A^2\frac{\alpha_i\alpha_j}{|\vec{\alpha}|^2(|\vec{\alpha}| + 1)}$$

A.2 Appendix A2

Here, we calculate the derivative of Φ with respect to Θ_j.

$$\begin{aligned}
\frac{\partial}{\partial\Theta_j}\Phi(\vec{X}, \Theta, \Lambda) &= \frac{\partial}{\partial\Theta_j}\sum_{i=1}^{N} ln(\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j)P(j)) \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial\Theta_j} ln(\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j)P(j)) \\
&= \sum_{i=1}^{N}\frac{1}{\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j)P(j)}\frac{\partial}{\partial\Theta_j}(\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j)P(j)) \\
&= \sum_{i=1}^{N}\frac{1}{p(\vec{X}_i, \Theta)}\frac{\partial}{\partial\Theta_j}(\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j)P(j))
\end{aligned}$$

Since:

$$p(j/\vec{X}_i, \Theta_j) = \frac{p(\vec{X}_i, \Theta_j)P(j)}{p(\vec{X}_i, \Theta)} \tag{56}$$

Then

$$\frac{\partial}{\partial\Theta_j}\Phi(\vec{X}, \Theta, \Lambda) = \sum_{i=1}^{N} p(j/\vec{X}_i, \Theta_j)\frac{\frac{\partial}{\partial\Theta_j}\sum_{j=1}^{M} p(\vec{X}_i/j, \Theta_j)P(j)}{p(\vec{X}_i/j, \Theta_j)P(j)} \tag{57}$$

which is equivalent to:

$$\frac{\partial}{\partial \Theta_j}\Phi(\vec{X},\Theta,\Lambda) = \sum_{i=1}^{N} p(j/\vec{X}_i,\Theta_j)\frac{\frac{\partial}{\partial \Theta_j}p(\vec{X}_i/j,\Theta_j)P(j)}{p(\vec{X}_i/j,\Theta_j)P(j)} \tag{58}$$

Finally, we obtain:

$$\frac{\partial}{\partial \Theta_j}\Phi(\vec{X},\Theta,\Lambda) = \sum_{i=1}^{N} p(j/\vec{X}_i,\Theta_j)\frac{\partial}{\partial \Theta_j} ln(p(\vec{X}_i/j,\Theta_j)P(j)) \tag{59}$$

A.3 Appendix A3

Here, we calculate the derivative of $ln(p(\vec{X}_i/j,\vec{\alpha_j}))$ with respect to α_{jl}. We have:

$$p(\vec{X}_i/j,\vec{\alpha_j})) = \frac{\Gamma(|\vec{\alpha_j}|)}{A^{|\vec{\alpha_j}|-1}\prod_{l=1}^{dim+1}\Gamma(\alpha_{jl})}\prod_{l=1}^{dim+1} X_{il}^{\alpha_{jl}-1}$$

Then

$$ln(p(\vec{X}_i/j,\vec{\alpha_j})) = ln(\Gamma(|\vec{\alpha_j}|)) - ln(A^{|\vec{\alpha_j}|-1}) - ln(\prod_{l=1}^{dim+1}\Gamma(\alpha_{jl})) + ln(\prod_{l=1}^{dim+1} X_{il}^{\alpha_{jl}-1})$$

So,

$$\begin{aligned}\frac{\partial}{\partial \alpha_{jl}} ln(p(\vec{X}_i/j,\vec{\alpha_j})) &= \frac{\partial}{\partial \alpha_{jl}} ln(\Gamma(|\vec{\alpha_j}|)) - \frac{\partial}{\partial \alpha_{jl}} ln(A^{|\vec{\alpha_j}|-1}) - \frac{\partial}{\partial \alpha_{jl}} ln(\prod_{l=1}^{dim+1}\Gamma(\alpha_{jl})) \\ &+ \frac{\partial}{\partial \alpha_{jl}} ln(\prod_{l=1}^{dim+1} X_{il}^{\alpha_{jl}-1})\end{aligned}$$

Since we have:

$$\frac{\partial}{\partial x} ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)} = \Psi(x) \tag{60}$$

Then:

$$\begin{aligned}\frac{\partial}{\partial \alpha_{jl}} ln(\prod_{l=1}^{dim+1}\Gamma(\alpha_{jl})) &= \frac{\partial}{\partial \alpha_{jl}}\sum_{l=1}^{dim+1} ln(\Gamma(\alpha_{jl})) \\ &= \frac{\partial}{\partial \alpha_{jl}} ln(\Gamma(\alpha_{jl})) \\ &= \Psi(\alpha_{jl})\end{aligned}$$

We have also:

$$\begin{aligned}\frac{\partial}{\partial\alpha_{jl}}ln(\prod_{l=1}^{dim+1}X_{il}^{\alpha_{jl}-1}) &= \frac{\partial}{\partial\alpha_{jl}}\sum_{l=1}^{dim+1}ln(X_{il}^{\alpha_{jl}-1}) \\ &= \frac{\partial}{\partial\alpha_{jl}}ln(X_{il}^{\alpha_{jl}-1}) \\ &= \frac{\partial}{\partial\alpha_{jl}}(\alpha_{jl}-1)ln(X_{il}) = ln(X_{il})\end{aligned}$$

and

$$\frac{\partial}{\partial\alpha_{jl}}ln(A^{|\vec{\alpha_j}|-1}) = \frac{\partial}{\partial\alpha_{jl}}(|\vec{\alpha_j}|-1)ln(A) = ln(A)$$

Finally we obtain:

$$\frac{\partial}{\partial\alpha_{jl}}ln(p(\vec{X}_i/j,\vec{\alpha_j})) = \Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A)$$

A.4 Appendix A4

Here, we calculate the derivative of Φ with respect to β_{jl}.

$$\begin{aligned}\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda) &= \frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda)\times\frac{\partial\alpha_{jl}}{\partial\beta_{jl}} \\ &= \frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda)\times\frac{\partial e^{\beta_{jl}}}{\partial\beta_{jl}} \\ &= \frac{\partial}{\partial\alpha_{jl}}\Phi(\vec{X},\Theta,\Lambda)\times e^{\beta_{jl}}\end{aligned}$$

Since $e^{\beta_{jl}} = \alpha_{jl}$, then:

$$\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda) = \alpha_{jl}[(\Psi(|\vec{\alpha_j}|)-\Psi(\alpha_{jl}))\sum_{i=1}^{N}p(j/\vec{X}_i,\vec{\alpha_j}) + \sum_{i=1}^{N}p(j/\vec{X}_i,\vec{\alpha_j})(ln(X_{il})-ln(A))]$$

A.5 Appendix A5

Here, we calculate the second and the mixing derivatives of the log-Likelihood function with respect to β_{jl}.

$$\frac{\partial^2}{\partial\beta_{jl}^2}\Phi(\vec{X},\Theta,\Lambda)$$

$$
\begin{aligned}
&= \frac{\partial}{\partial\beta_{jl}}(\frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda)) \\
&= \frac{\partial}{\partial\beta_{jl}}[e^{\beta_{jl}} \times [(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) + \sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il}) - ln(A))]] \\
&= [e^{\beta_{jl}} \times [(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) + \sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il}) - ln(A))]] \\
&\quad +[e^{\beta_{jl}} * [e^{\beta_{jl}} \times (\Psi^{'}(|\vec{\alpha_j}|) - \Psi^{'}(\alpha_{jl}))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) + (\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}))\sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl}} p(j/\vec{X_i},\vec{\alpha_j}) \\
&\quad + \sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl}} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il}) - ln(A))]] \\
&= \frac{\partial}{\partial\beta_{jl}}\Phi(\vec{X},\Theta,\Lambda) + \alpha_{jl}^2(\Psi^{'}(|\vec{\alpha_j}|) - \Psi^{'}(\alpha_{jl}))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) \\
&\quad +\alpha_{jl}[\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}))\sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl}} p(j/\vec{X_i},\vec{\alpha_j}) \\
&\quad + \sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl}} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il}) - ln(A))]
\end{aligned}
$$

$\Psi^{'}(.)$ is the Trigamma function and we have:

$$\Psi^{'}(x) = \frac{\partial}{\partial x}\Psi(x) = \frac{\partial^2}{\partial x^2} ln(\Gamma(x)) \tag{61}$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial\beta_{jl_1}\partial\beta_{jl_2}}\Phi(\vec{X},\Theta,\Lambda) \\
&= \frac{\partial}{\partial\beta_{jl_1}}(\frac{\partial}{\partial\beta_{jl_2}}\Phi(\vec{X},\Theta,\Lambda)) \\
&= \frac{\partial}{\partial\beta_{jl_2}}[e^{\beta_{jl_1}} \times [(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl_1}))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) + \sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il_1}) - ln(A))]] \\
&= [e^{\beta_{jl_1}} \times [e^{\beta_{jl_2}} \times (\Psi^{'}(|\vec{\alpha_j}|))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) + (\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl_1}))\sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl_2}} p(j/\vec{X_i},\vec{\alpha_j}) \\
&\quad + \sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl_2}} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il_1}) - ln(A))]] \\
&= \alpha_{jl_1}\alpha_{jl_2}\Psi^{'}(|\vec{\alpha_j}|))\sum_{i=1}^{N} p(j/\vec{X_i},\vec{\alpha_j}) + \alpha_{jl_1}[\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl_1}))\sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl_2}} p(j/\vec{X_i},\vec{\alpha_j}) \\
&\quad + \sum_{i=1}^{N} \frac{\partial}{\partial\beta_{jl_2}} p(j/\vec{X_i},\vec{\alpha_j})(ln(X_{il_1}) - ln(A))]
\end{aligned}
$$

A.6 Appendix A6

Here, we calculate the derivative of $p(j/\vec{X}_i, \vec{\alpha_j})$ with respect to β_{jl}. We know that:

$$\frac{\partial}{\partial \beta_{jl}} p(j/\vec{X}_i, \vec{\alpha_j}) = \alpha_{jl} \frac{\partial}{\partial \alpha_{jl}} p(j/\vec{X}_i, \vec{\alpha_j})$$

Since:

$$\begin{aligned}
& \frac{\partial}{\partial \alpha_{jl}} p(j/\vec{X}_i, \vec{\alpha_j}) \\
= & \frac{\partial}{\partial \alpha_{jl}} \frac{p(\vec{X}_i/j, \vec{\alpha_j})P(j)}{\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j)} \\
= & \frac{P(j)\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j}) \sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j) - P(j)p(\vec{X}_i/j, \vec{\alpha_j})\frac{\partial}{\partial \alpha_{jl}} \sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j)}{(\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j))^2} \\
= & \frac{P(j)\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j}) \sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j) - P(j)p(\vec{X}_i/j, \vec{\alpha_j})\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j})P(j)}{(\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j))^2} \\
= & \frac{P(j)\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j})(\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j) - P(j)p(\vec{X}_i/j, \vec{\alpha_j}))}{(\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j))^2}
\end{aligned}$$

Using the following result (see Appendix A7):

$$\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j}) = p(\vec{X}_i/j, \vec{\alpha_j})(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A)) \tag{62}$$

we obtain:

$$\begin{aligned}
& \frac{\partial}{\partial \beta_{jl}} p(j/\vec{X}_i, \vec{\alpha_j}) \\
= & \alpha_{jl}(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A))[\frac{P(j)p(\vec{X}_i/j, \vec{\alpha_j})}{\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j)} - \frac{(P(j)p(\vec{X}_i/j, \vec{\alpha_j}))^2}{(\sum_{j=1}^{M} p(\vec{X}_i/j, \vec{\alpha_j})P(j))^2}] \\
= & \alpha_{jl}(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A))(p(j/\vec{X}_i, \vec{\alpha_j}) - p(j/\vec{X}_i, \vec{\alpha_j})^2) \\
= & \alpha_{jl}(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A))p(j/\vec{X}_i, \vec{\alpha_j})(1 - p(j/\vec{X}_i, \vec{\alpha_j}))
\end{aligned}$$

A.7 Appendix A7

Here, we calculate the derivative of $p(\vec{X}_i/j, \vec{\alpha_j})$ with respect to β_{jl}. For this purpose, we use Eq. 36. We have:

$$\frac{\partial}{\partial \alpha_{jl}} ln(p(\vec{X}_i/j, \vec{\alpha_j})) = \Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A))$$

Since:

$$\frac{\partial}{\partial \alpha_{jl}} ln(p(\vec{X}_i/j, \vec{\alpha_j})) = \frac{\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j})}{p(\vec{X}_i/j, \vec{\alpha_j})}$$

Thus

$$\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j}) = p(\vec{X}_i/j, \vec{\alpha_j}) \frac{\partial}{\partial \alpha_{jl}} ln(p(\vec{X}_i/j, \vec{\alpha_j}))$$

Finally:

$$\frac{\partial}{\partial \alpha_{jl}} p(\vec{X}_i/j, \vec{\alpha_j}) = p(\vec{X}_i/j, \vec{\alpha_j}) \frac{\partial}{\partial \alpha_{jl}} (\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl}) + ln(X_{il}) - ln(A))$$

B Appendix B

Experiments have shown that the following approximations can be made without affecting the final results:

$$(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl})) \sum_{i=1}^{N} \frac{\partial}{\partial \beta_{jl}} p(j/\vec{X}_i, \vec{\alpha_j}) + \sum_{i=1}^{N} \frac{\partial}{\partial \beta_{jl}} p(j/\vec{X}_i, \vec{\alpha_j}) (ln(X_{il}) - ln(A)) \simeq 0$$

Thus, Eq. 39 becomes:

$$\frac{\partial^2}{\partial^2 \beta_{jl}} \Phi(\vec{X}, \Theta, \Lambda) = \frac{\partial}{\partial \beta_{jl}} \Phi(\vec{X}, \Theta, \Lambda) + \alpha_{jl}^2 (\Psi^{'}(|\vec{\alpha_j}|) - \Psi^{'}(\alpha_{jl})) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) \tag{63}$$

Besides, we have noted that:

$$(\Psi(|\vec{\alpha_j}|) - \Psi(\alpha_{jl_1})) \sum_{i=1}^{N} \frac{\partial}{\partial \beta_{jl_2}} p(j/\vec{X}_i, \vec{\alpha_j}) + \sum_{i=1}^{N} \frac{\partial}{\partial \beta_{jl_2}} p(j/\vec{X}_i, \vec{\alpha_j}) (ln(X_{il_1}) - ln(A)) \simeq 0$$

Thus, Eq. 40 becomes:

$$\frac{\partial^2}{\partial \beta_{jl_1} \partial \beta_{jl_2}} \Phi(\vec{X}, \Theta, \Lambda) = \alpha_{jl_1} \alpha_{jl_2} \Psi^{'}(|\vec{\alpha_j}|) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) \tag{64}$$

According to Eq. 43 the information matrix I is:

$$I_{l_1 l_2} = -\alpha_{jl_1} \alpha_{jl_2} \Psi^{'}(|\vec{\alpha_j}|) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}), \quad l_1 \neq l_2 \tag{65}$$

$$I_{l_1 l_1} = -\frac{\partial}{\partial \beta_{jl_1}} \Phi(\vec{X}, \Theta, \Lambda) - \alpha_{jl_1}^2 (\Psi'(|\vec{\alpha_j}|) - \Psi'(\alpha_{jl_1})) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) \tag{66}$$

This matrix can be written as:

$$I = D + \delta a b^T, \tag{67}$$

where

$$D = diag[D_1, \ldots, D_{dim+1}] \tag{68}$$

$$D_l = -\frac{\partial}{\partial \beta_{jl}} \Phi(\vec{X}, \Theta, \Lambda) + \alpha_{jl}^2 \Psi'(\alpha_{jl}) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) \tag{69}$$

$$\delta = -\Psi'(|\vec{\alpha_j}|) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) \tag{70}$$

$$a^T = b^T = (\alpha_{j1}, \ldots, \alpha_{jdim+1}) \tag{71}$$

The variance-covariance matrix is obtained as the inverse of the Fisher's information matrix I by a well-known theorem (Theorem 8.3.3) given by Graybill [17]. The variance-covariance matrix $V = v_{l_1 l_2}$ is thus found to be:

$$V = I^{-1} = D^* + \delta^* a^* a^{*T}, \tag{72}$$

where:

$$D^* = diag[1/D_1, \ldots, 1/D_{dim+1}] \tag{73}$$

$$a^{*T} = (\alpha_{j1}/D_1, \ldots, \alpha_{jdim+1}/D_{dim+1}) \tag{74}$$

$$\delta^* = \Psi'(|\vec{\alpha_j}|) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j})(1 - \Psi'(|\vec{\alpha_j}|) \sum_{i=1}^{N} p(j/\vec{X}_i, \vec{\alpha_j}) \sum_{l=1}^{dim+1} \frac{\alpha_{jl}^2}{D_l}) \tag{75}$$

References

[1] Amari, S. Information Geometry of the EM and em Algorithm for Neural Networks. *Neural Networks*, 8(9):1379–1408, 1995.

[2] Amari, S. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10:251–276, 1998.

[3] Amari, S., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L. and Rao, C. R. *Differential Geometry in Statistical inference*, volume 10. Institute of Mathematical Statistics, Hayward, California, 1987.

[4] Beckman, R. J. and Tietjen, G. L. Maximum Likelihood Estimation for the Beta Distribution. *Journal of Statistical Computation and Simulation*, 7:253–258, 1978.

[5] Betchel, Y. C., Bonaiti-Pelli, C., Poisson, N., Magnette, J. and Bechtel, P. R. A Population and Family Study of N-acetyltransferase Using Caffeine Urinary Metabolites. *Clinical Pharmacology and Therapeutics*, 54:134–141, 1993.

[6] Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.

[7] Bishop, C. M. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[8] Choi, S. C. and Wette, R. Maximum Likelihood Estimation of Parameters of the Gamma Distribution and Their Bias. *Technometrics*, 11(4):683–690, November 1969.

[9] Crawford, S. L. An Application of the Laplace Method to Finite Mixture Distributions. *Journal of the American Statistical Association*, 89:259–267, 1994.

[10] Dempster, A. P., Laird, N. M. and Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.

[11] Duda, R. O. and Hart, P. E. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[12] El Zaart, A. and Ziou, D. Statistical Modeling of Multimodal SAR Images. *To appear in the International Journal of Remote Sensing*, 2002.

[13] Everitt, B. S. and Hand, D. J. *Finite mixture Distributions.* Chapman and Hall, London, UK., 1981.

[14] Fielitz, B. D and Myers, B. L. Estimation of Parameters in the Beta Distribution. *Decision Sciences*, 6:1–13, 1975.

[15] Genovese, C. and Wasserman, L. Rates of Convergence for the Gaussian Mixture Sieve. *The Annals of Statistics*, 28(4):1105–1127, 2000.

[16] Gnanadesikan, R., Pinkham, R. S. and Hughes, L. P. Maximum Likelihood Estimation of the Beta Distribution from Smallest Order Statistics. *Technometrics*, 9(4):607–620, 1967.

[17] Graybill, F. A. *Matrices with applications in Statistics.* Wadsworth, California, 1983.

[18] Ikeda, S. Acceleration of the EM algorithm. *Systems and Computers in Japan*, 31(2):10–18, February 2000.

[19] Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data.* John Wiley, New York, 1990.

[20] Kherfi, M. L., Ziou, D. and Bernardi, A. Content-Based Image Retrieval Using Positive and Negative Examples. *To appear*, 2002.

[21] R. Kothari and D. Pitts. On Finding the Number of Clusters. *Pattern Recognition Letters*, 20:405–416, 1999.

[22] Kotz, S. and Ng, K. W. and Fang, K. *Symmetric Multivariate and Related Distributions.* London/New York: Chapman and Hall, 1990.

[23] Kotz, S., Balakrishman and Norman, J. *Continuous Multivariate Distributions*, volume 1. New York: Wiley-Interscience, 2000.

[24] Medasani, S. and Krishnapuram, R. A Comparison of Gaussian and Pearson Mixture Modeling for Pattern Recognition and Computer Vision Applications. *Pattern Recognition Letters*, 20:305–313, 1999.

[25] Medasani, S. and Krishnapuram, R. Categorization of Image Databases for Efficient Retrieval Using Robust Mixture Decomposition. *Computer Vision and Image Understanding*, 83:216–235, 2001.

[26] Merz, C. J. and Murphy, P. M. UCI Repository of Machine Learning Databases. *[http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.*, 1996.

[27] Narayanan, A. A Note on Parameter Estimation in the Multivariate Beta Distribution. *Computer Mathematics and Applications*, 24(10):11–17, 1992.

[28] Newsman, S., Sumengen, B. and Manjunath, B. S. Category-Based Image Retrieval. In *Proc. IEEE International Conference on Image Processing, Special Session on Multimedia Indexing, Browsing and Retrieval*, Thessalonica, Greece, September 2001.

[29] Nguyen, T. B. and Ziou, D. Contextual and Non-Contextual Performance Evaluation of Edge Detectors. *Pattern Recognition Letters*, (21):805–816, 2000.

[30] Ortiz, L. E. and Kaelbling, L. P. Notes on Methods Based on Maximum-Likelihood Estimation for Learning the Parameters of the Mixture of Gaussian Model. Technical Report CS-99-03, Computer Science Department, Brown University, February 28 1999.

[31] Raftery, A. E. and Banfield, J. D. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49:803–821, 1993.

[32] Rao, C. R. *Advanced Statistical Methods in Biomedical Research.* New York: John Wiley and Sons, 1952.

[33] Rao, P. and B. L. S. *Asymtotic Theory of Statistical Inference.* Wiley Series in Probability and Mathematical Statistics, 1987.

[34] Redner, R. A. and Walker, H. F. Mixture Densities, Maximum Likelihood and EM Algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[35] Roberts, S. J. Parametric and Non-Parametric Unsupervised Cluster Analysis. *Pattern Recognition*, 30(2):261–272, 1997.

[36] Roberts, S. J. and Rezek, L. Bayesian Approach to Gaussian Mixture Modeling. *Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, 1998.

[37] Ronning, G. Maximum Likelihood Estimation of Dirichlet Distributions. *Journal of Statistical Computation and Simulation*, 32:215–221, 1989.

[38] Santner,T. J. and Duffy, D. E. *The Statistical Analysis of Discrete Data.* Springer-Verlag, 1989.

[39] Weis, G. H. and Dishon, M. Small Sample Comparison of Estimation Methods for the Beta Distribution. *Journal of Statistical Computation and Simulation*, 11:1–11, 1980.

[40] Xu, L. and Jordan, M. I. On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, 8(1):129–151, 1996.

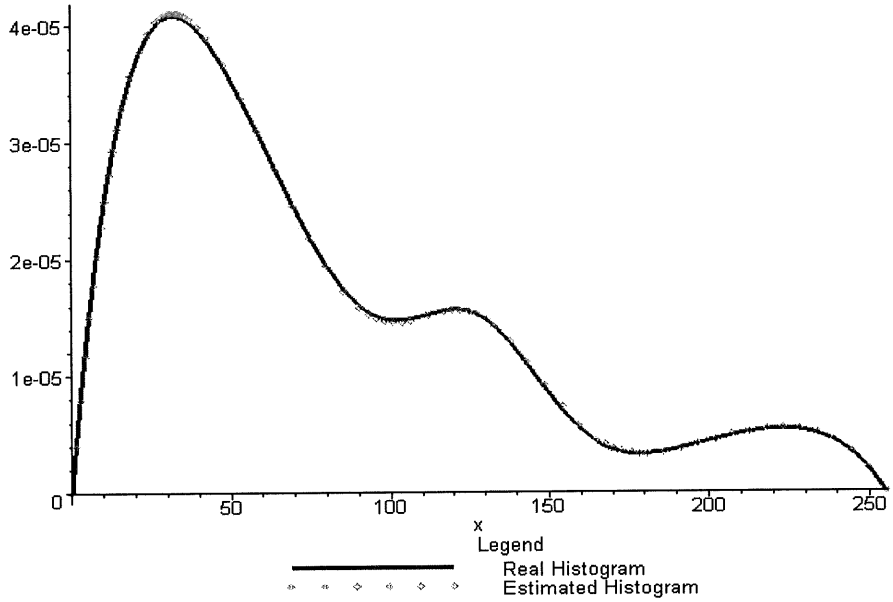


Figure 5: Real and estimated histograms of the third synthetic data set.

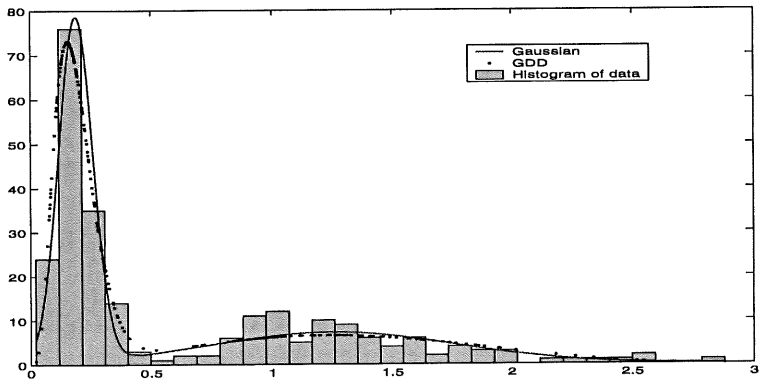


Figure 6: Real and estimated histograms for the Enzyme data set.

	Real parameters	Estimated parameters
Mode 1	$P(1)$=0.75	$P(1)$=0.746
	$\alpha_{11} = 2$	$\alpha_{11} = 2.043$
	$\alpha_{12} = 8$	$\alpha_{12} = 8.307$
Mode 2	$P(2)$=0.15	$P(2)$=0.156
	$\alpha_{21} = 20$	$\alpha_{21} = 19.185$
	$\alpha_{22} = 20$	α_{22} =19.086
Mode 3	$P(3)$=0.1	$P(3)$=0.098
	$\alpha_{31} = 8$	$\alpha_{31} = 8.220$
	$\alpha_{32} = 2$	α_{32} =2.035

Table 3: Estimation of the parameters of the GDD for the third synthetic data set.

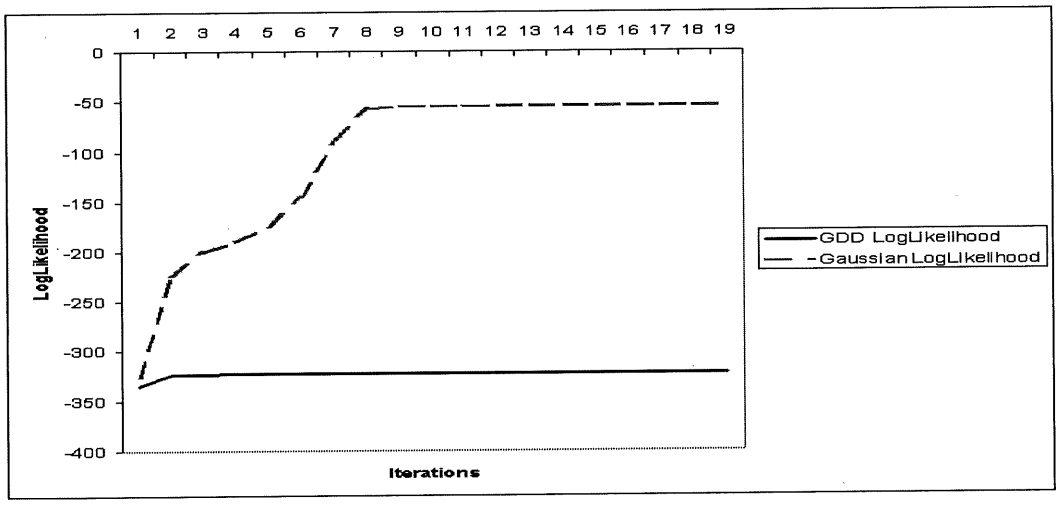


Figure 7: The likelihood cycle in the case of the Gaussian and the GDD for the Enzyme data set.

	GDD Mixture		Gaussian Mixture	
Estimated parameters	class1	class2	class1	class2
	$P(1)$=0.396	$P(2)$=0.604	$P(1)$=0.408	$P(2)$=0.592
	α_{11}=3.201	α_{21}=4.894	μ_1=1.253	μ_2=0.187
	α_{12}=4.185	α_{22}=72.401	σ_1^2=0.263	σ_2^2=0.005
Number of iterations	9		17	

Table 4: Estimation of the parameters of the GDD and Gaussian mixtures for the Enzyme data set.

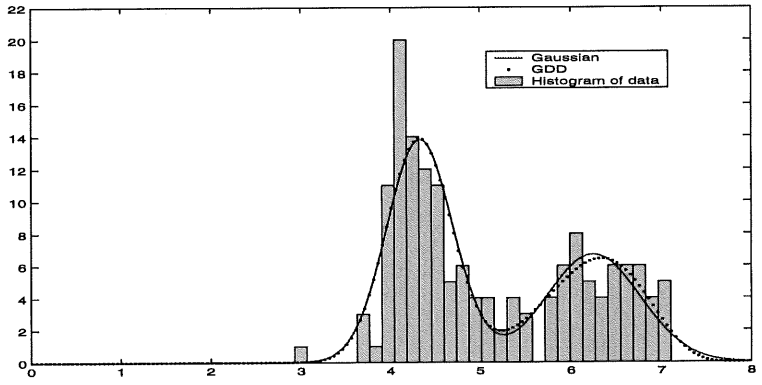


Figure 8: Real and estimated histograms for the acidity data set.

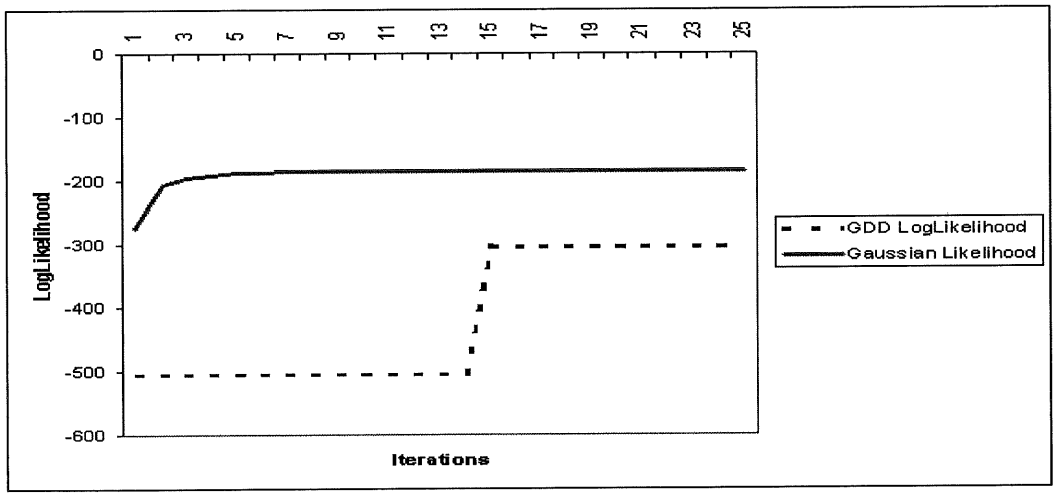


Figure 9: The likelihood cycle in the case of the Gaussian and the GDD for the Acidity data set.

	GDD Mixture		Gaussian Mixture	
Estimated parameters	class1	class2	class1	class2
	$P(1)$=0.579	$P(2)$=0.421	$P(1)$=0.596	$P(2)$=0.404
	α_{11}=66.570	α_{21}=26.656	μ_1=4.330	μ_2=6.249
	α_{12}=56.969	α_{22}=7.760	σ_1^2=0.138	σ_2^2=0.270
Number of iterations	14		20	

Table 5: Estimation of the parameters of the GDD and Gaussian mixtures for the Acidity data set.

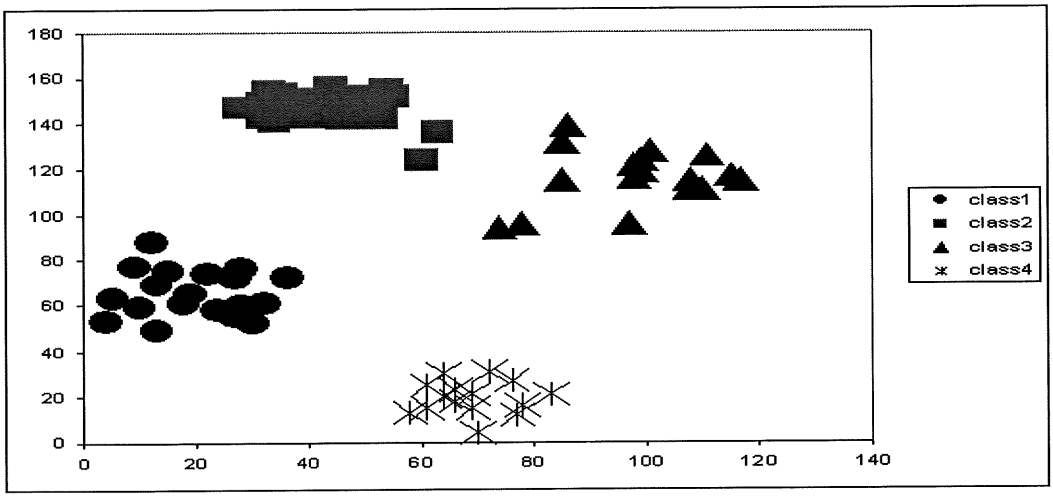


Figure 10: The Ruspini data set.

	GDD Mixture			
Parameters	class1	class2	class3	class4
	$P(1)$=0.267	$P(2)$=0.295	$P(3)$=0.239	$P(4)$=0.199
	α_{11}=4.709	α_{21}=23.023	α_{31}=15.465	α_{41}=29.289
	α_{12}=16.121	α_{22}=78.146	α_{32}=18.781	α_{42}=7.914
	α_{13}=40.487	α_{23}=32.315	α_{33}=5.744	α_{43}=68.230
Error	%1.33			
Iterations	10			

Table 6: Estimation of the parameters of the GDD mixture for the Ruspini data set.

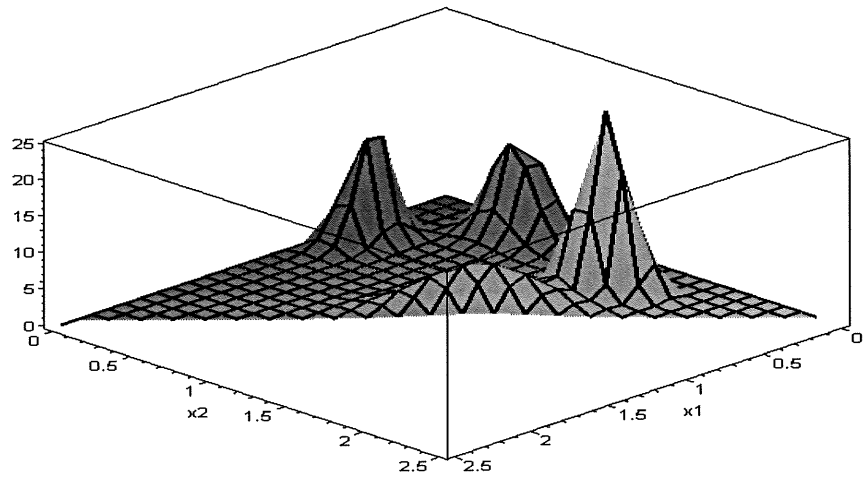


Figure 11: Representation of the Ruspini data set by a GDD mixture.

	Gaussian Mixture			
	class1	class2	class3	class4
Parameters	$P(1)$=0.266	$P(2)$=0.294	$P(3)$=0.240	$P(4)$=0.2
	μ_{11}=20.15	μ_{21}=43.453	μ_{31}=95.720	μ_{41}=68.933
	μ_{12}=64.95	μ_{22}=147.046	μ_{32}=115.039	μ_{42}=19.4
	$V1=\begin{pmatrix} 44.013 & -2.396 \\ -2.396 & 48.223 \end{pmatrix}$	$V2=\begin{pmatrix} 43.481 & -6.348 \\ -6.348 & 14.752 \end{pmatrix}$	$V3=\begin{pmatrix} 77.426 & 30.245 \\ 30.245 & 56.640 \end{pmatrix}$	$V4=\begin{pmatrix} 23.897 & 0.113 \\ 0.113 & 24.653 \end{pmatrix}$
Error	2.66 %			
Iterations	11			

Table 7: Estimation of the parameters of the Gaussian mixture for the Ruspini data set.

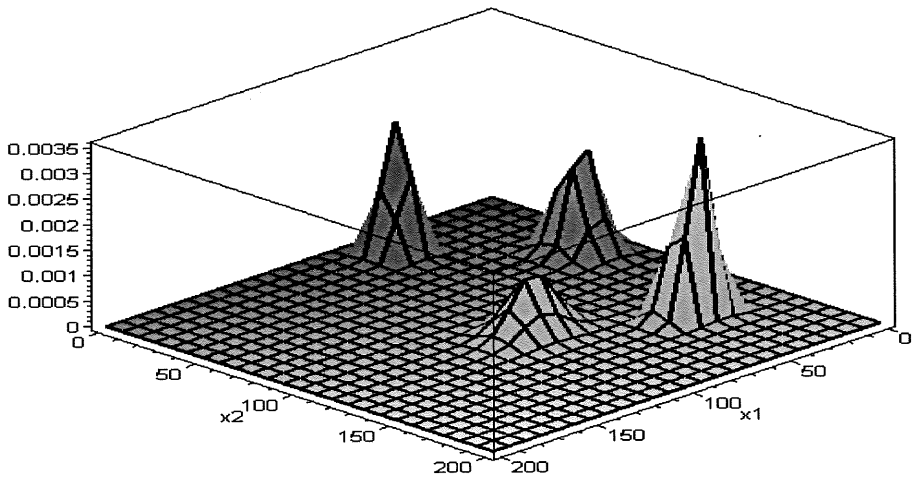


Figure 12: Representation of the Ruspini data set by a Gaussian mixture.

Gaussian Mixture

$$v1=\begin{pmatrix} 2.164e-5 & -6.973e-6 & 3.684e-7 & 1.103e-7 & 4.083e-6 & -4.025e-6 & 4.607e-6 & -3.948e-6 & -7.473e-8 \\ -6.973e-6 & 1.820e-5 & 2.282e-7 & 8.014e-7 & -3.327e-6 & 4.249e-6 & -3.994e-6 & 3.906e-6 & 1.640e-7 \\ 3.684e-7 & 2.282e-7 & 7.775e-7 & 1.632e-7 & 2.745e-7 & -1.283e-7 & 2.675e-7 & -1.540e-8 & 3.417e-8 \\ 1.103e-7 & 8.014e-7 & 1.632e-7 & 1.138e-6 & 7.958e-8 & 7.272e-8 & 4.159e-8 & 1.045e-7 & 3.222e-8 \\ 4.083e-6 & -3.327e-6 & 2.745e-7 & 7.958e-8 & 2.982e-6 & -2.015e-6 & 2.167e-6 & -2.152e-6 & -3.044e-9 \\ -4.025e-6 & 4.249e-6 & -1.283e-7 & 7.272e-8 & -2.015e-6 & 3.105e-6 & -2.251e-6 & 2.275e-6 & 4.563e-8 \\ 4.607e-6 & -3.994e-6 & 2.675e-7 & 4.156e-8 & 2.167e-6 & -2.251e-6 & 6.399e-6 & -2.236e-6 & -1.301e-8 \\ -3.948e-6 & 3.906e-6 & -1.540e-8 & 1.045e-7 & -2.152e-6 & 2.275e-6 & -2.236e-6 & 5.802e-6 & 7.884e-8 \\ -7.473e-8 & 1.640e-7 & 3.417e-8 & 3.222e-8 & -3.044e-9 & 4.563e-8 & -1.301e-8 & 7.884e-8 & 9.288e-8 \end{pmatrix}$$

μ_1=(0.019 0.017 0.010 0.010 0.015 0.014 0.015 0.014 0.010)

$P(1)$=0.442

$$v2=\begin{pmatrix} 0.0001 & 2.282e-5 & 3.634e-5 & 1.633e-5 & 2.726e-5 & 4.772e-5 & 2.350e-5 & 2.271e-5 & -7.793e-6 \\ 2.282e-5 & 0.0001 & 7.0755e-5 & 5.415e-5 & 4.251e-5 & 4.387e-5 & 5.821e-5 & 4.662e-5 & 3.821e-5 \\ 3.634e-5 & 7.075e-5 & 0.0001 & 7.516e-5 & 6.139e-5 & 6.816e-5 & 6.607e-5 & 6.099e-5 & 5.289e-5 \\ 1.633e-5 & 5.415e-5 & 7.516e-5 & 0.0001 & 4.772e-5 & 5.243e-5 & 6.224e-5 & 5.049e-5 & 5.697e-5 \\ 2.726e-5 & 4.251e-5 & 6.139e-5 & 4.772e-5 & 9.523e-5 & 4.288e-5 & 6.037e-5 & 4.374e-5 & 4.135e-5 \\ 4.772e-5 & 4.387e-5 & 6.816e-5 & 5.243e-5 & 4.288e-5 & 0.0001 & 5.431e-5 & 4.333e-5 & 3.134e-5 \\ 2.350e-5 & 5.821e-5 & 6.607e-5 & 6.224e-5 & 6.037e-5 & 5.431e-5 & 0.0001 & 5.552e-5 & 4.655e-5 \\ 2.271e-5 & 4.662e-5 & 6.099e-5 & 5.049e-5 & 4.374e-5 & 4.333e-5 & 5.552e-5 & 9.872e-5 & 4.825e-5 \\ -7.793e-6 & 3.821e-5 & 5.289e-5 & 5.697e-5 & 4.135e-5 & 3.134e-5 & 4.655e-5 & 4.825e-5 & 0.0001 \end{pmatrix}$$

μ_2=(0.039 0.053 0.048 0.045 0.042 0.049 0.049 0.044 0.033)

$P(2)$=0.558

Error= 2.342%

Table 8: Estimation of the parameters of the Gaussian mixture for the Breast Cancer data set.

	GDD Mixture	
	class1	class2
Parameters	$P(1)$=0.625	$P(2)$=0.375
	α_{11}=8.595	α_{21}=3.482
	α_{12}=4.323	α_{22}=3.113
	α_{13}=4.556	α_{23}=3.146
	α_{14}=4.444	α_{24}=2.457
	α_{15}=7.079	α_{25}=2.668
	α_{16}=4.283	α_{26}=3.374
	α_{17}=6.604	α_{27}=2.949
	α_{18}=4.156	α_{28}=2.547
	α_{19}=3.983	α_{29}=1.273
	α_{110}=292.75	α_{210}=21.617
Error	1.024 %	

Table 9: Estimation of the parameters of the GDD mixture for the Breast Cancer data set.

	Class1	Class2	Class3	Class4	Class5
Class1	101	0	10	0	0
Class2	0	120	0	13	0
Class3	0	0	108	0	0
Class4	0	6	0	104	4
Class5	0	2	0	5	127

Table 10: Confusion matrix for image classification by a GDD mixture.

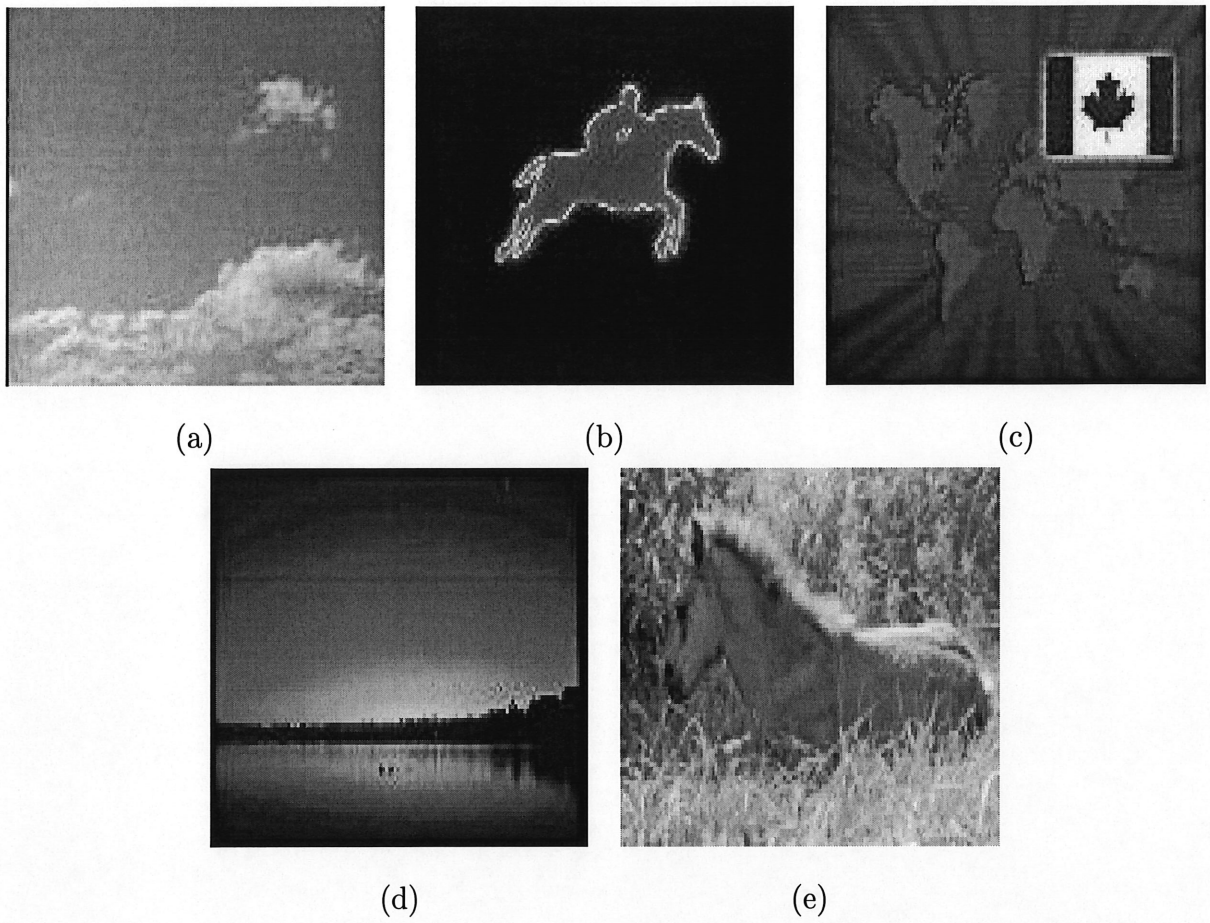


Figure 13: Sample images from each group. (a) Class1, (b) Class2, (c) Class3, (d) Class4, (e) Class5.

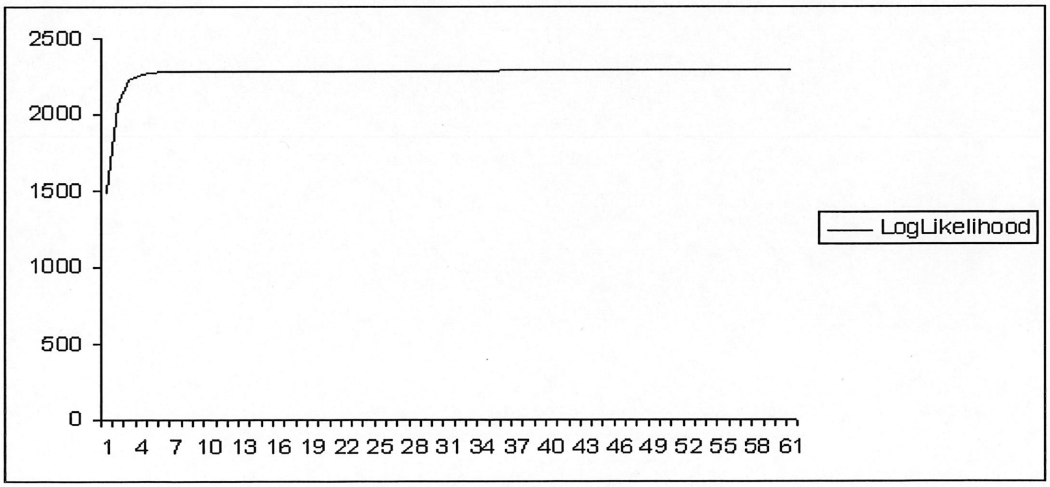


Figure 14: The likelihood cycle in the case of the image-summarizing application.

Conclusion

Dans ce mémoire, nous nous sommes intéressés aux problèmes relatifs à la recherche d'images, particulièrement médicales. L'étude détaillée de ces problèmes nous a permis de comprendre les éléments clés de ces systèmes.

Dans la première partie du travail, nous avons étudié chacun des problèmes relatifs à la recherche d'images médicales afin de comprendre les spécificités de ce domaine. D'abord il y a l'identification et l'extraction des caractéristiques et des métadonnées qui décrivent le mieux les images. Ensuite, il y a la définition des mesures de similarité qui correspondent mieux à l'usager. Nous avons étudié également le problème de l'indexation et de la catégorisation de la base de données d'images et nous pensons que c'est un point très important dans la recherche d'images. Cette étude a été suivie par une analyse des systèmes existants pour la recherche d'images médicales. Nous pensons que la recherche d'images médicales est un domaine en pleine expansion et de grande complexité à cause de la quantité d'information produite par la médecine et la diversité des domaines.

Comme nous pensons que les mixtures sont un moyen très efficace pour l'indexation et pour résumer les bases de données nous nous sommes focalisés sur ce sujet dans la deuxième partie du mémoire. Grâce aux mixtures, les données peuvent être partitionnées en catégories homogènes et ainsi la recherche sera plus facile et plus rapide. Contrairement aux travaux classiques qui utilisent la loi normale comme densité, nous avons utilisé la distribution de Dirichlet qui présente plusieurs avantages. Notre méthode a été validée par plusieurs évaluations non-contextuelles basées sur des histogrammes synthétisés et contextuelles concernant la classification des données et le résumé des bases de données d'images. Le travail que nous avons présenté est une première étape. D'autres questions restent à être abordées telle que la détermination automatique du nombre des classes.

Bibliographie

[1] Bouguila, N., Ziou, D. and Vaillancourt, J. A Maximum Likelihood Estimation of the Generalized Dirichlet Mixture. *Submitted.*

[2] Bouguila, N., Ziou, D. and Vaillancourt, J. An Overview of Medical Image Retrieval Systems. Technical report, DMI, faculté des sciences, Université de Sherbrooke, 2002.

[3] Bouguila, N., Ziou, D. and Vaillancourt, J. The introduction of Dirichlet Mixtures into Computer Vision and Pattern Recognition Applications. *Submitted to Computer Vision and Pattern Recognition*, June 16-22 2003.

[4] Eakins, J. P. and Graham, M. E. Content-based Image Retrieval. A report to the JISC Technology Applications Program. Technical report, Institute for Images Data Research, University of Northumbria at Newcastle, January 1999.

[5] El Zaart, A. and Ziou, D. Statistical Modeling of Multimodal SAR Images. *To appear in the International Journal of Remote Sensing*, 2002.

[6] Kherfi, M. L., Ziou, D. and Bernardi, A. Web Images Search Engines : A survey. Technical Report 276, DMI, faculté des sciences, Université de Sherbrooke , December 2001.

[7] Rui, Y. Image Retrieval : Current Techniques, Promising Directions, and Open Issues. *Journal of visual communication and image representation*, 10(4) :39–62, April 1999.

[8] Rui, Y. and Huang, T. S. Image retrieval : Past, Present, and future. In *International symposium on multimedia information*, 1997.

[9] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R. Content-Based Image Retrieval at the End of the Early Years. *Pattern Recognition and Machine Intelligence*, 22(12) :1349–1380, December 2000.

[10] Veltkamp, R. C. and Tanase, M. Content-Based Image Retrieval Systems. Technical Report UU-CS-2000-34, Department of Computing Science, Utrecht University , March 2001.