Chapter 1

Introduction

1.1 Information Retrieval

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources[1]. The inception of information retrieval can be traced back to the early 1960s. A search engine is the practical application of information retrieval techniques to large-scale text collections. Web search engines are the most common applications of the information retrieval systems. Other applications are visible at many universities and public libraries where they provide access to books, journals and other documents by using some information retrieval systems. Over the years, Information Retrieval commonly referred to as IR, has evolved and adapted to suit the various information

[1] https://en.wikipedia.org/wiki/Information_retrieval

needs ranging from simple users, corporations to professionals and scientists. When distinguished by the scale they operate at, IR systems can be broadly classified into three types [61]:

- Web search
- Enterprise, institutional and domain-specific search
- Personal information retrieval

An information retrieval process begins when a user enters a query into the system. The queries are formal definitions of information needs of the users. In commonplace, the query is a string of words through which search process is done by some web search engine or IRs [50]. Such query based search process does not fetch a uniquely identified object. Instead, it provides a list of matching objects from the collection. Those retrieved objects may not be equally relevant to the given query. While traditional IR began with finding useful information from *unstructured* text data, it quickly evolved to encompass various multimedia elements like images, audio, video *etc.* with the advent of the Internet.

1.2 Image Retrieval

Image Retrieval is concerned with searching, browsing and retrieval from a large database of digital images. Generally, images are searched against any user-provided

query terms such as keywords or clicking on some images. As such they are also a form of information retrieval where the information is disseminated as images. Image search engines like Bing[2], Yahoo[3] or online photo-sharing like Flickr[4] *etc.* return images similar to the given query. Based on the search mode, conventional Image Retrieval systems are broadly subdivided into two categories:

1. Text-Based Image Retrieval System

2. Content-Based Image Retrieval System.

In both cases, initially, a matching score is calculated between query (irrespective of its *mode*) and images, present in the dataset. Based on the matching scores images relevant images are ranked and retrieved.

1.2.1 Text Based Image Retrieval

In these systems, text retrieval techniques are used on textual annotations or text descriptions of images to perform image retrieval. Generally, the images are manually annotated by text descriptors from the available textual contents present in the database [93]. The available textual contents can be anything like keywords, subject headings, captions, or natural language text *etc.* In such cases, user provided queries are textual *i.e.* keywords and these keywords are matched with the available image annotations to retrieve relevant images.

[2]https://www.bing.com/
[3]https://in.yahoo.com/
[4] https://www.flickr.com/

- **Advantages:** Textual descriptions are easier to process.

- **Drawbacks:** Manually annotation of images needs a considerable amount of human intervention and this human labor is expensive as well as time-consuming. The cost associated with manual annotation is prohibitive with regards to a large-scale data set. Besides, these manual annotations are not always reliable due to the subjectivity of human perception. Also very often, the visual content of images can hardly be described in the text. Thus the human-provided textual descriptions provide very little information about the image contents. Moreover, in the real world, it is hard to expect that all images will be uniformly annotated. Frequently annotations are missing or may contain noise. So, such text-only based retrieval methods suffer when annotations are missing or contain noisy annotations.

To overcome the issues mentioned above, Content-Based Image Retrieval was introduced as an alternative to Text-based image retrieval and has gradually gained momentum.

1.2.2 Content Based Image Retrieval

Conventional Content-Based Image Retrieval (CBIR) systems rely on extracting individual features and process them, *i.e.*, they involve extracting visual features from the query image and perform a search on those feature plane on the database. Here, the query images as well as all the target images, present in the search space,

are represented as a set/collection of features. The relevance of a query image and a target image is calculated based on some similarity scores. Generally, the low-level features like color, shape, texture *etc.*, are used for comparison [37]. Semantic features such as the type of object present in the image are also used although these semantic features are hard to extract and thus it remains an open problem despite several attempts to address it.

Drawbacks Although content-based image retrieval overcomes the constraints of text-based image retrieval, it suffers from some disadvantages [72] which we elaborate as follows:

- **The Image Domain:** A narrow image search domain has a limited and predictable variability whereas a broad image search domain has an unlimited and unpredictable variability. However, in the current era of humongous online data and storage, image search domain is always enormous. In reality, broad domain images are polysemic and their semantics may be available partially. Also, the interpretation of any recorded scene may not be unique. So, content-based image search has become more and more complex with the passage of time.

- **The Sensory Gap:** As it is very hard to describe exactly any object in machine level, there is always a gap between the object and information from its computational description. It yields to uncertainty about the real world

object and it affects mostly when precise recording conditions are missing. So there occurs a sensory gap.

- **The Semantic Gap:** Linguistic description of an image is nearly impossible. So a gap exists between the extracted information from visual data and its interpretation. Although in content-based image retrieval, image descriptions rely on data-driven features, these features are disconnected from the human interpretation. Hence there is always a semantic gap.

- **The curse of dimensionality** Conventional content-based image retrieval systems classifies images into different categories to reduce the search space. But, sometimes they may not be helpful to retrieval as an image may belong to more than one category and it creates confusion. To ease the problem, frequently the number of features is increased. However, this increased number of features is detrimental to the cause.

As both the text-based and content-based image retrieval systems suffer from various limitations, the research community has gradually pivoted towards a new approach called multimodal image retrieval.

1.2.3 Multimodal Image Retrieval

[5]https://blogs.ischool.utexas.edu/perspectives/files/2013/04/Screen-Shot-2013-04-12-at-1.24.41-PM.png

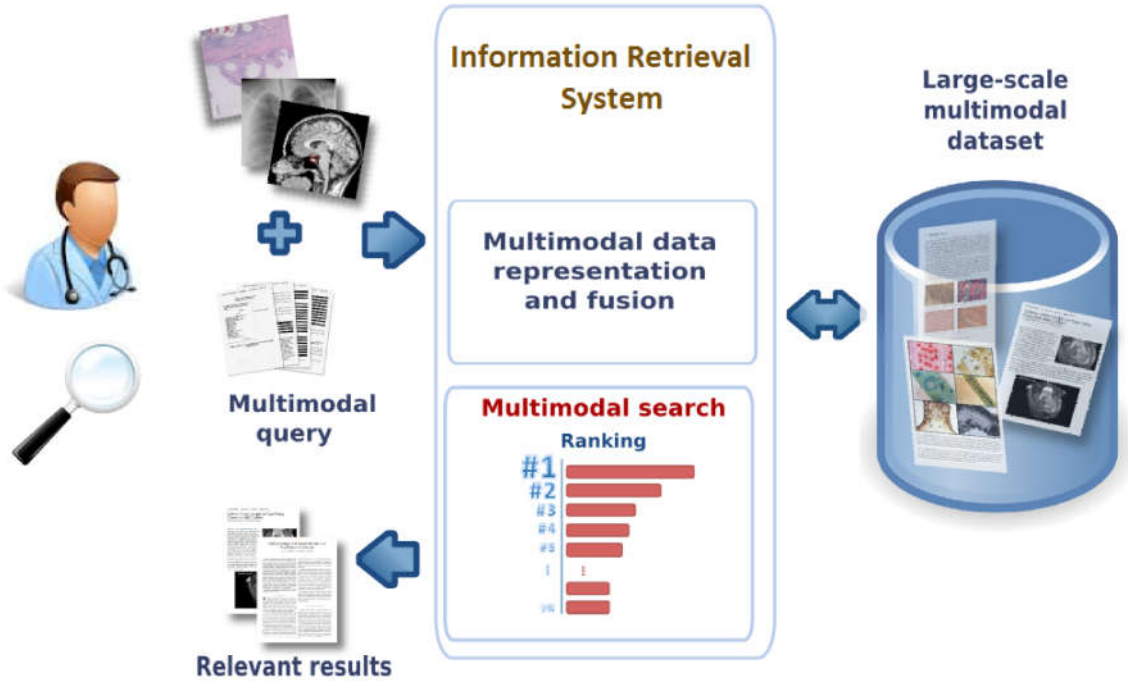

FIGURE 1.1: A diagrammatic overview of a multimodal IR system[5]

To bridge the semantic gap between the high-level information need of users and commonly employed low-level features, Multimodal Image Retrieval has become an inevitable solution and a new trend of research. Information from different sources like text, images, video *etc.*, called *modes*, are combined together and then the infused information is used to retrieve images (and sometimes text) efficiently. Since image retrieval is also a form of information retrieval, the phrases "multimodal image retrieval" and "multimodal IR" are used interchangeably. In multimodal IR system, queries are the combination of different types of data instead of a simple text string. The multimodal queries are infused information needs which are expressed in terms of a combination of regular texts, images, audios *etc.*

Figure 1.1 depicts a diagrammatic overview of a basic multimodal IR system. Like any other IR system, a multimodal object is an entity that is represented by different

sources of information in a content collection or database. User provided queries are matched against the database or collection of information, and finally, the retrieved results are typically ranked according to their relevancy. Unlike classical SQL queries of a database, in IR systems, the retrieved results may or may not be exact match to the given query.

In this thesis, we consider text and image as different modalities. However, establishing the relationship between image content and accompanying text description not a trivial task. For different feature combination fusion is a well-established technique. Generally, two types of fusion technique or strategy are available in the literature: (*a*) Early fusion and (*b*) Late fusion [83]. In early fusion, extracted features are combined first and then the combined information is sent to a single analysis unit which generates the decision. Whereas in late fusion, first the local decisions are taken based on individual features and then all the local decisions are combined. Finally, the combined decision is analyzed and a final decision is obtained based on that. Feature fusion is described in more details in Section 1.2.4.4.

1.2.4 Preliminaries

In this section we briefly discuss some prerequisites which should help in understanding the content of the thesis better. Few key terms and phrases are introduced in brief as follows.

1.2.4.1 Relevance Feedback

The idea behind the relevance feedback is to involve the user in the retrieval process to improve the retrieval performance[6]. Generally, in an interacting framework, the users provide some feedback against the retrieved results regarding how much the results are relevant to their needs or not. The user judgment against the initial retrieval phase is then incorporated into the next phase of retrieval process to improve the retrieval efficiency. Usually, the system generates a better representation of the information need based on the user provided feedback or in other words, the initial query is upgraded based on the user feedback. Finally, the system returns a revised set of retrieved results which are supposed to be more relevant than the initial set of results. There are mainly three types of relevance feedback: (*a*) Explicit feedback, (*b*) Implicit feedback and (*c*) Blind or pseudo feedback feedback[7].

- *Explicit feedback*: When user-provided feedback or judgments are available from an interactive system, and these judgments are directly incorporated for revised retrieval process, it is known as explicit feedback [14]. In such cases, users mark the relevance explicitly using a binary or graded relevance system. In binary relevance system, the documents are marked as either relevant or not relevant whereas in graded relevance system indicates the relevance of a document to a query on a scale using numbers, letters, or descriptions (such as "not relevant", "somewhat relevant", "relevant", or "very relevant" *etc*).

[6]https://nlp.stanford.edu/IR-book/html/htmledition/relevance-feedback-and-pseudo-relevance-feedback-1.html

[7]https://en.wikipedia.org/wiki/Relevance_feedback

- *Implicit feedback*: Implicit feedback is inferred from user behavior, such as noting which documents they do and do not select for viewing, the duration of time spent viewing a document, or page browsing or scrolling actions[8]. Here the users are not directly providing any judgment or feedback instead of the system continuously monitors the user behaviour and collects information regarding user needs. Based on this observation, the system modifies the base query and fetches a new set of results. That is why it is called implicit feedback.

- *Blind feedback*: Pseudo or blind relevance feedback is a method where the system automatically analyzes the retrieved documents and re-retrieves more relevant document without any external intervention. In this method, the system selects top k number of documents from an initial ranked list of documents and treat them as relevant. After analyzing these top-ranked documents, the system infers some knowledge about user needs and accordingly modifies the search process. As the entire process is performed without any user intervention or user knowledge, the process is called blind feedback.

1.2.4.2 Query Expansion

Query expansion is one of the most common techniques to reformulate a seed query or user given query. This process involves evaluating the context of a given query or user's information needs. Based on the additional information obtained through relevance feedback or some other sources the original query is modified. Often,

[8] www.scils.rutgers.edu/etc/mongrel/kelly-belkin-SIGIR2001.pdf

relevant terms or keyphrases, obtained from analyzing the relevant documents or from user's behavior, are appended with the original query to improve the retrieval performance. When direct user interaction is not available, thesaurus-based query expansion is used and for such cases, direct user input is not required. For such systems, a thesaurus is formed from different sources like (a) a controlled vocabulary which is maintained by human editors, (b) a manual thesaurus, (c) an automatically derived thesaurus based on co-occurrence statistics over a collection of documents or (d) thesaurus formed by query log mining *etc.*

1.2.4.3 Keyphrase Extraction

Keyphrase extraction is commonly known as the extraction of important topical words and phrases from the body of a document. These extracted keyphrases or terms are considered important since they provide a concise description of the document. Whereas a keyphrase is a sequence of one or more words that are considered highly relevant, a keyword is a single word that is highly relevant which can be used as the basis for finding candidate phrases. Thus by identifying the keywords, keyphrases can be identified. Usually, keyphrases are more informative than keywords. There are various applications of keyphrase extraction such as document categorization, clustering, indexing, search, summarization, quantifying semantic similarity with other documents *etc.*

Any typical keyphrase extraction system operates in two phrases: (1) Extracting a list of words or phrases that serve as candidate keyphrases (potential keyphrases)

using some heuristics and (2) determining which of these candidate keyphrases are correct using either some supervised or unsupervised approaches [39]. In recent times, the unsupervised approaches have gained much success rather than the supervised approaches in terms of performance.

1.2.4.4 Fusion

Information regarding the same object can be obtained from different sources. The nature of the acquired information or data varies depending on the data acquisition processes. Each data type (out of various types of data) is called a modality. The increasing availability of multiple modalities of data regarding a particular object provides better analysis than what any single mode of data can provide. So, there is an increasing trend of combining the multiple modalities of data. This growing need leads us to data fusion. Data fusion can be defined as an analysis of several datasets such that different datasets can interact and inform each other [49]. However, data fusion is not a trivial task as the data are generated from complex systems, due to diversity and heterogeneity of datasets. Many fusion strategies are available in the literature and those are broadly categorized into three types. Among them, the widely used strategy is to fuse the information at the feature level, which is also known as early fusion. The other approach is decision level fusion or late fusion which fuses multiple modalities in the semantic space [4]. A combination of these approaches is also practiced as the hybrid fusion approach.

1.2.5 Measuring Metrics

The key to the evaluation measures of any information retrieval system hinges on how well the retrieved results satisfy the user's query intent. In such an IR framework, the instances are the documents and the task is to return a set of relevant documents against the search query. Eventually, all such documents can be categorized into two classes. Those documents which are relevant to the search query are treated as 'relevant' set and the rest which are treated to be 'not relevant' set of documents. Based on the relevancy of retrieved result, the performance of the system is adjudged. This notion of relevancy is known as the ground truth or gold standard. We choose a few among those measuring metrics such as recall, precision, MAP *etc.* to adjudge our models. A brief description of these metrics is provided in the following subsections.

1.2.5.1 Recall

The number of relevant items retrieved by the system is quantitatively measured by Recall. From information retrieval point of view, recall is the fraction of the relevant documents out of all the documents that are successfully retrieved [79]. From statistical viewpoint, it can be defined as the probability that a certain relevant document will be retrieved by a search process. For better understanding, we can

define recall by the following Equation 1.1

$$recall = \frac{|(relevant\ documents) \cup (retrieved\ documents)|}{|(relevant\ documents)|} \tag{1.1}$$

Achieving recall of 100% is very difficult since no systems are perfect. So, we cannot simply rely only on recall value to judge a system. Hence, measuring the precision value becomes crucial to judge a system's accuracy correctly.

1.2.5.2 Precision

In layman's words, precision can be defined as the measure of how many retrieved documents are relevant. The fraction of the retrieved documents that are relevant to a given query is called precision in information retrieval [6]. It can be defined by the Equation 1.2.

$$precision = \frac{|(relevant\ documents) \cup (retrieved\ documents)|}{|(retrieved\ documents)|} \tag{1.2}$$

Precision can also be termed as the probability that a certain retrieved document is relevant to the given query. While calculating the precision value, all retrieved documents can be considered but also it can be evaluated at a given cut-off rank [61]. In such case, the top-most results, retrieved by a system, are taken into account and such type of measure is called "Precision at k" or $P@k$.

1.2.5.3 F-measure

The combined effect of precision and recall is measured by a metric called F-measure. It is the inverse of the mean value of the precision and recall and can be expressed by the following Equation 1.3:

$$Fmeasure = 2 \cdot \frac{recall \cdot precision}{recall + precision} \tag{1.3}$$

Sometimes it is also called as the harmonic average of the precision and recall. For perfect accuracy *i.e.* the best value of F-measure can be 1 while the worst is 0.

1.2.5.4 MAP

Mean Average Precision (MAP) is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved against a set of query and this value is then averaged over information needs *i.e.* the given set of query [61]. The Equation 1.4 refers the MAP.

$$MAP = \frac{\sum_{q=1}^{Q} AvgP(q)}{Q} \tag{1.4}$$

where Q is the total number of query present and $AvgP(q)$ is the per-query average precision. MAP can be treated as a single-figure measure of retrieval quality. Among evaluation measures of information retrieval, MAP has the best stability and good

amount of discrimination. This is one of the reasons why most of the research work in this thesis has been evaluated using MAP.

1.3 Thesis Motivation and Objective

In present times, the enormous volume of ever-growing web data is making the task of retrieval tougher day by day. Under such circumstances, single modality based retrieval systems like text-only based or image-only based system, often fails to achieve optimal performance. Hence, multimodal retrieval has attracted a lot of attention in image retrieval field over the past few years. Continuing the trend, this thesis too considers textual and visual data as two different modes for multimodal tasks. The reason behind is that generally textual descriptions are easier to process and more comprehensible for user but at the same time, it is very hard to describe any visual content *e.g.* images, through text. So the entire thesis work incorporates both the text and the image as two different sources of information. Each such modality consists of a set of features. Conventionally, the visual features extracted from images are treated as low-level features whereas the textual features are regarded as high level features. Sometimes different features are extracted from a single modality, such as, image features like interior border, color, texture *etc.* are extracted as local features, whereas shape, contour *etc.* are extracted as global features [56].

Once modalities are determined and features extracted, the next step is how to combine these different modalities such that best possible retrieval performance can be

obtained. To accomplish the feature combination, mainly two types of feature fusion strategies are found in the literature; Early fusion and Late Fusion [5]. Determining not only the fusion strategy but also assigning the proper weight for each feature to combine is crucial for multimodal information search. Often, for proper weight learning an intelligent decision-making process is adapted to favour the feature set according to their importance.

In any information retrieval system, query is the fundamental substance to express the information needs of users and formulation of query is of prime importance. Also, an insightful query formulation leads towards an efficient retrieval result. However, it is not always easy for any naive user to express her/his information needs in a machine comprehensible way, reformulation of any existing or user-provided query is pivotal. Being motivated by this fact, this thesis uses Query Expansion using relevant documents, a widely accepted approach, to reformulate query. Pseudo-Relevance Feedback[9] is used for query expansion in all of our works that involve query expansion. For doing so, appropriate terms are deduced from the relevant documents and ranked according to how useful the terms might be for fetching more relevant documents. From the ranked list of informative terms, few top-ranked terms are then appended with the original query to make it more comprehensible for the retrieval system. To deduce important terms from relevant documents we employ Keyphrase Extraction techniques and thus we study the effect of query reformulation in the multimodal image retrieval. In this context we study and compare few well-established keyphrase extraction techniques and also propose a novel one.

[9]http://nlp.stanford.edu/IR-book/html/htmledition/pseudo-relevance-feedback-1.html

A good indexing and orientation of data is also a key to better retrieval efficiency and without annotation image indexing is not feasible. Hence, our next step of research is towards image annotation. We propose a novel automatic image annotation model employing two different statistical measure; co-occurrence count and mutual information. The novelty of this work lies in capturing the image semantic at entire image level and also not losing the information shared at the object level. The wholesome information which is conveyed by the entire scene is depicted as 'concept' and the acquired knowledge from each discrete object are the elements of the concept. A community detection technique is applied to aid our model and to improve the annotation performance. To establish the superiority of our proposed model, it is compared with other existing systems.

The semantic gap between different modalities like text and image is still an open challenge and multimodal solutions are being enthusiastically studied to mitigate the gap. In light of this, other than the image retrieval, we verify the multimodal solution on another task that is image classification. The main crux of multimodal image classification is how to effectively combine the various features to improve the performance. For this we improvise a well established algorithm *i.e.* Hill Climbing on a support vector based multiple kernel classification framework. It is a well known fact in artificial intelligence that hill climbing algorithm suffers from local optimization problem. To overcome this, we propose an improved version of hill climbing algorithm and we name it as "Extended Hill Climbing". Our proposed combination approach outperforms the other state-of-the-art techniques. The experiments are

carried out on a standard dataset to validate our claim.

1.4 Contribution

The contribution of this thesis is many-fold:

1. Different keyphrase extraction techniques are available in the literature of text retrieval. However, to the best of our knowledge, this is the first attempt which studies the effects of keyphrase extraction based query reformulation on multimodal image retrieval.

2. We also hypothesize that inclusion of relevant part of the narratives of a text query may significantly enhance the image retrieval efficiency. This need to stress on relevant part stems from our objective to bring the user closer to her needs in an intelligent fashion.

3. We propose a new keyphrase extraction approach that tries to capture the semantic relationship between words apart from their co-occurrence. This is actualized with the aid of a word graph formed using mutual information and WordNet. These semantically enriched keyphrases are then used for textual query expansion. In other words, we try to minimize the semantic gap that exists between the image and text by intelligently processing the associated text. The selection of keyphrases from the word graph is made through a greedy algorithm.

4. We adopt a topic model based keyphrase extraction technique to compare in our study. In this modal, Latent Dirichlet Allocation (LDA) is used to build a Topical PageRank (TPR) on word graph to measure word importance with respect to different topics.

5. Our next contribution is to annotate images automatically. We present a novel graph-based approach for automatic image annotation to describe the complex scenes where there is a possibility of the presence of multiple objects and the image in itself conveys a particular concept. We measure the relatedness between concepts in two ways; the first one is by using co-occurrence count and the second one is by using mutual information. First of all, we construct a graph comprising of all the concepts termed as '*concept-graph*'. On this concept graph, we apply an information theoretic approach for community detection using Map Equation to detect and form clusters of similar or overlapping concepts. Finally, we traverse through the identified cluster of concepts using a greedy walk to obtain new annotations for the test image or to refine the available annotations. To the best of our knowledge, this is the first attempt at employing mutual information for capturing informativeness of concepts towards enhancing automatic image annotation.

6. Not only is our image annotation task is novel but also our proposed approaches outperform several baseline methods by a significant margin. We validate our claim through a set of experiments on a publicly available standard dataset (ImageCLEF 2012 Scalable image annotation task) and finally observe that

the mutual information based approach supersedes the co-occurrence based approach.

7. We hypothesize that multimodal solutions may work better in any other task like retrieval. So we study the effect of multimodal solution on image classification task too. One vital aspect of any such multimodal task is to combine the involved features appropriately as an optimal result can be obtained efficiently. In light of the arguments provided above, we propose a technique to optimize the weight of each feature (both image and text) and also compute the optimal combined feature weights. We employ Hill Climbing, a well-established optimization technique, for feature optimization. Many works in the literature have employed various artificial intelligence based methods for feature selection and combination. However, we believe that our work is the first attempt that utilizes Hill Climbing strategy towards multimodal feature combination.

8. Since Hill Climbing suffers from local optimization, a new improved variation of Hill Climbing is proposed to overcome the constrained Hill Climbing, we call the improved version as Extended Hill Climbing. Finally, using the optimally combined feature set, we perform image classification on an SVM framework. Our proposed classification method outperforms both the average combination and the weighted average combination based classification. An exhaustive set of experiments is carried out on a standard real-world dataset to validate our claim.

1.5 Thesis Organization

The rest of this thesis is organized as follows: **Chapter 2** discusses few of the related works which have in someway inspired, influenced and possibly paved the way for the research work done for this thesis. **Chapter 3** introduces the problem of multimodal retrieval, points out the possible research directions and finally proposes a graph-based keyphrase extraction technique exploiting the term relations through statistical measures for reformulating user query for better performance in text-image retrieval. **Chapter 4** further revises and improves the technique proposed in Chapter 3 by introducing a Topic model based technique for keyphrase extraction, which is believed to be the first work in this direction for multimodal problems. Automatic image annotation, as explained in previous sections is a hot research topic, but there is scope for improvement as is the case for most research problems. **Chapter 5** explores this problem and suggests a solution to this using concepts of community detection. **Chapter 6** deals with another serious issue of multimodality *i.e.* classification and aims to solve it by proposing an extension to a well established artificial intelligence technique. A conclusion to the dissertation is drawn in **Chapter 7**.