A Survey of Indexing and Retrieval of Multimodal Documents: Text and Images

Technical Report 2006-505

Nawei Chen
chenn@cs.queensu.ca
School of Computing
Queen's University
Kingston, Ontario, Canada

February 2006

Abstract

A document conveys information using multiple modalities, including text, layout/style and images. For example, journal articles usually have figures to illustrate experimental results, and the title in a journal article usually has a different font size than the body text. Indexing and retrieval using only text is the traditional way of IR (Information Retrieval). With the development of the Internet and Digital Libraries, it becomes increasingly important to develop IR techniques for intelligent indexing and retrieval of multimodal documents, such as web pages in HTML or XML format, scientific publications in PDF format and document images from scanned papers. In this paper, I make a survey of multimodal IR systems that combine the text and image modalities. Indexing and retrieval are two important components of an IR system. Given a collection of documents, indexing describes documents using an index language. Retrieval uses the results of indexing and finds related documents corresponding to a user's query. Text and image modalities use different indexing and retrieval techniques. Single-modality IR, either using text or images, has limitations. Multimodal IR aims to overcome the limitations in each single modality by combining them. The following issues in multimodal IR are addressed: various techniques to combine text and images; techniques to find relationships between text and images; noise and uncertainties in IR systems; and techniques to improve effectiveness of IR, such as Latent Semantic Indexing, user's relevance feedback, semantic network, and document clustering and classification.

Keywords: multimodal Information Retrieval, multimodal documents, document image retrieval, CBIR, text retrieval

Table of Content

1. Introduction

Information Retrieval (IR) has different scopes of definitions. Broadly, IR is defined as "the field concerned with the structure, analysis, organization, storage, searching and retrieval of information" [Salt89]. Narrowly, IR refers to *ad hoc retrieval*, which is the task of "searching a collection of documents for relevant documents given an information need" [YaNe99]. In ad hoc retrieval, no prior knowledge is available on the relevance of documents. The system can only use the description of the information needs, which is often called a *query*. Typical examples of ad hoc retrieval systems are the search engines on the web, e.g. Google (www.google.com). In this survey, IR represents ad hoc retrieval.

Indexing and retrieval are two important components of an Information Retrieval system. Given a collection of documents, *indexing* describes documents using an index language [Rijs79]. *Retrieval* uses the results of indexing and finds related documents corresponding to a user's query. In this survey, we focus on indexing and retrieval techniques for multimodal documents. In this section, first we introduce multimodal documents, and then we provide an overview of Information Retrieval systems. In the end, some issues are raised to guide our survey.

1.1. Multimodal documents

A document conveys information using multiple modalities, including text, layout/style and images. For example, journal articles usually have figures to illustrate experimental results, and the title in a journal article usually has a different font size than the body text. Examples of multimodal documents are shown in Figure 1. Indexing and retrieval using only text is the traditional way of IR. With the development of the Internet and Digital libraries, it becomes increasingly important to develop IR techniques for intelligent indexing and retrieval of multimodal documents.

Figure 1. Examples of multimodal documents

There is a great diversity of computer-accessible multimodal documents. Based on the source of creation, there are two major categories, *electronic documents* and *document images*. *Electronic documents* are generated from computational document editing tools. Examples include web pages in HTML or XML format, and scientific publications in PDF format. *Document images* (also called document page images) are scanned from paper documents or from fax machine, and are represented in image format (TIFF, JPEG, etc). A document image corresponds to a single page in a multi-page document. We will survey IR techniques for both types of multimodal documents.

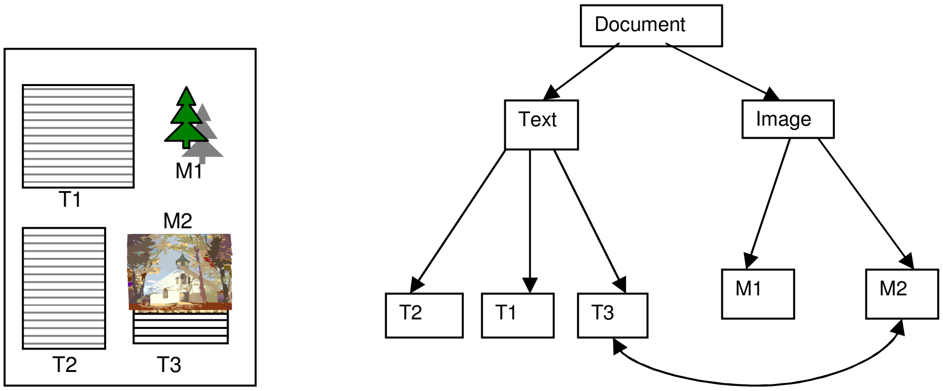


Figure 2. A multimodal document model

We define a multimodal document model as a set of multiple objects with two object types: text (T) or image (M). A text object represents a block of text. An image object represents an image. Images include graphical images (e.g. charts, graphs, maps, diagrams) and natural images (e.g. color or gray-scale photographs). Some text blocks have no association with images; these are called *independent text*. Other text blocks are associated with images, such as captions, or a set of keywords annotated to the images; these are called *collateral text* in this survey. Some images may not have collateral text. Figure 2 shows an example of a multimodal document. Text block T3 is related to the image M2. T1 and T2 are independent text blocks. Image M1 doesn't have collateral text. The right figure shows the hierarchical structure of the example multimodal document and the relationship between objects. Samples in Figure 1 can be represented in such a model. Most of the multimodal document models in our survey are simplified versions of the model in Figure 2, with each multimodal document containing a single image and possibly a collateral text block.

Unlike in electronic documents, there are no explicit objects in document images, so processing of document images relies on document image analysis techniques to find the text and image objects. A simplified diagram of document image analysis is shown in Figure 3. At first, text and images are separated from the document image through page segmentation and layout analysis. In this stage, collateral text may be associated with images. To get textual content, Optical Character Recognition (OCR) is applied to the text blocks. Noise and errors may occur in document image analysis. These have to be considered in subsequent indexing and retrieval.

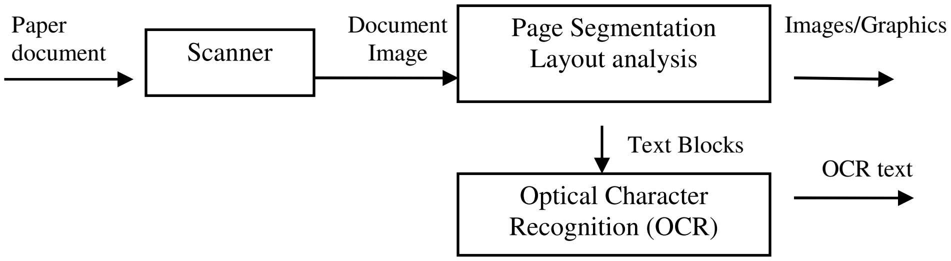


Figure 3. Procedures of document image analysis.

1.2. Overview of IR systems

Traditionally, a single modality, either text or images, has been used to retrieve content in multimodal documents. A simplified diagrammatic view of a single-modality IR system is shown in Figure 4. In general, there are four components in an IR system:

- **Indexing**: Indexing uses characteristic features to represent documents. Different features are extracted from either textual content of text blocks or visual content of images in a document depending on which modality is used. The indexing here refers to automatic indexing, i.e. indexes are automatically built without human intervention. The ideal indexing is to dynamically choose a set of features to represent documents given user's information needs. Automatic indexing techniques are surveyed in Section 2.
- **Query Formulating and Analyzing**: A user formulates a query through the query interface provided by the system. The system analyzes the query and represents it in the same internal format as used for document representations. Different systems differ in their friendliness and complexity of query interfaces depending on which modality is used for retrieval. A user query may be formulated in different ways. The query interface is important for users to form queries to represent their information needs. The details of query formulation and query interface are outside the scope of this paper.
- **Retrieval**: The system compares document representations and a query representation to retrieve documents using various retrieval models. Retrieval models are surveyed in Section 3. The result of a search is a set of *hits* containing both relevant (positive) documents and irrelevant (negative) documents.
- **Performance evaluation**: Precision and recall are the two most popular metrics to evaluate the effectiveness of text retrieval. *Precision* is the proportion of retrieved documents that are relevant, and *recall* is the proportion of relevant documents that are retrieved. Image retrieval and multimodal IR borrow these two terms for effectiveness evaluation. Techniques used for performance evaluation and benchmarking in IR are outside the scope of this paper.

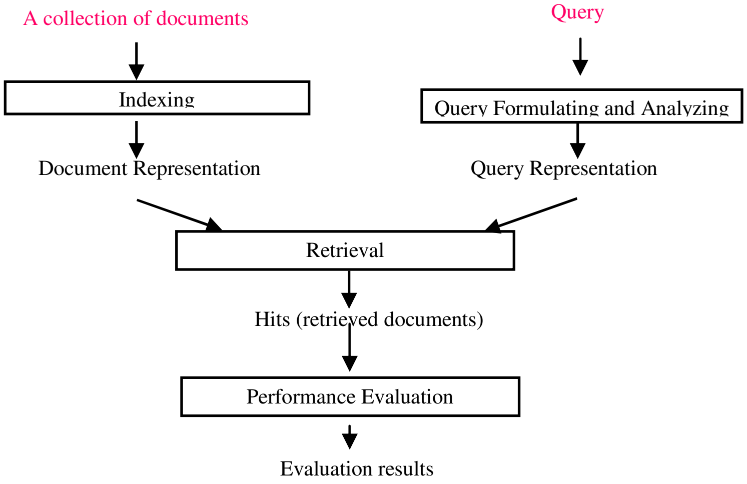


Figure 4. Diagrammatic view of a simplified single-modality IR system.

In the rest of this paper, we distinguish five types of IR systems: text retrieval, text-based image retrieval, CBIR (Content-Based Image Retrieval), document image retrieval and multimodal IR. They are briefly introduced in the following.

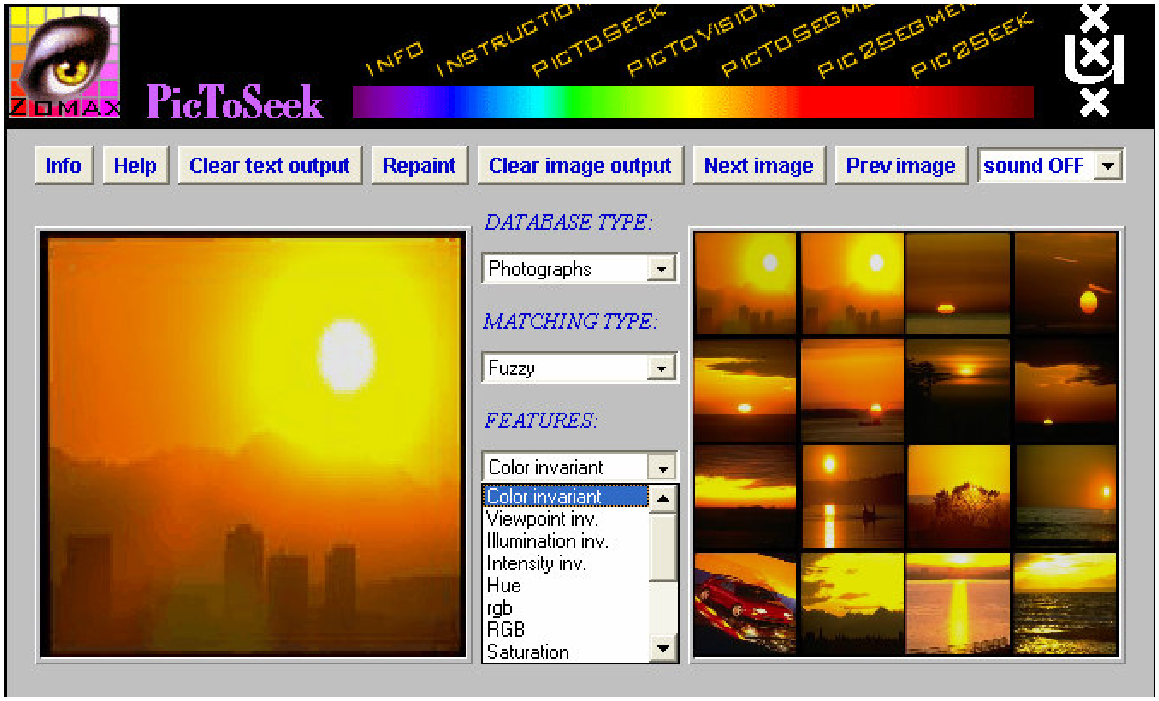
Text retrieval is the oldest branch of IR and it is well researched [Rijs79] [Salt89] [YaNe99] [Witt98]. Text retrieval only deals with text blocks, e.g. T1, T2, or T3 in Figure 2. A user represents a query using either a sequence of words or a Boolean combination *(AND, OR and NOT)* of words. For example, a user might input "*((image OR graphics) AND retrieval)*" or "image graphics retrieval" to find the information on image/graphics retrieval depending on the implemented retrieval model in the system. The indexing techniques are surveyed in Section 2.1.1 and retrieval models are surveyed in Section 3.1.

Text-based image retrieval and CBIR are two major approaches to retrieve images [Good00]. Text-based image retrieval is cross-modality retrieval, i.e. using collateral text to retrieve images. An example is to use collateral text block T3 to retrieve image M2 in Figure 2. The user represents a query using a textual string, and collateral text blocks are indexed. Collateral text blocks are linked to the image. There are uncertainties in finding collateral text blocks (discussed in Section 2.1.2). The retrieval techniques are similar to text retrieval or standard database retrieval using SQL [RuHC97]. We won't survey standard database retrieval techniques, which mostly deal with structured data.

CBIR deals with only images, e.g. M1 or M2 in Figure 2. CBIR uses visual content extracted directly from images for indexing and retrieving. The query may be an example image, a user's sketch or image attributes. An example of using an example image as a query is shown in Figure 5. An example of using a sketch or specifying color attributes as a query is shown in Figure 6. CBIR has drawn a lot of attention since the early 90's. There are several comprehensive surveys on CBIR [RuHC97] [RuHu99] [YoIc99] [SmWS00] [AnKJ02] [VeTa02]. Commercial CBIR systems are available, such as IBM's QBIC. Most of the systems surveyed in this paper are experimental systems in academia. Visual content indexing techniques are surveyed in Section 2.2. CBIR retrieval models are surveyed in Section 3.2.

Document image retrieval uses the results from page segmentation and/or document layout analysis (shown in Figure 3). After a preprocessing step to separate text from images, most document image retrieval systems deal with either text or images. There are comprehensive surveys on indexing and retrieval techniques for document images [Doer98] [MiCh00]. Textual content indexing and retrieval is similar to the techniques in text retrieval. However, errors in OCR must be considered. Basic textual content indexing methods are surveyed in Section 2.1.3. Visual content indexing is similar to the indexing in CBIR. The difference is that visual content may come from images within a document image or from the whole page image.

Multimodal IR systems use both the text and image objects. IR using either the text or image modality has advantages and limitations (surveyed in Section 4.1). The combination of text with images is demonstrated to improve retrieval performance [ScCS99] [BaFo01]. As far as we know, there is no existing survey of this field. Multimodal IR systems are based on single-modality IR systems and they are more complex than single-modality IR systems. Composite queries including both text and images may be used. Both textual and visual content are indexed and various combination techniques are developed to combine retrieval results using both text and images. We will survey combination of text and images in Section 4.

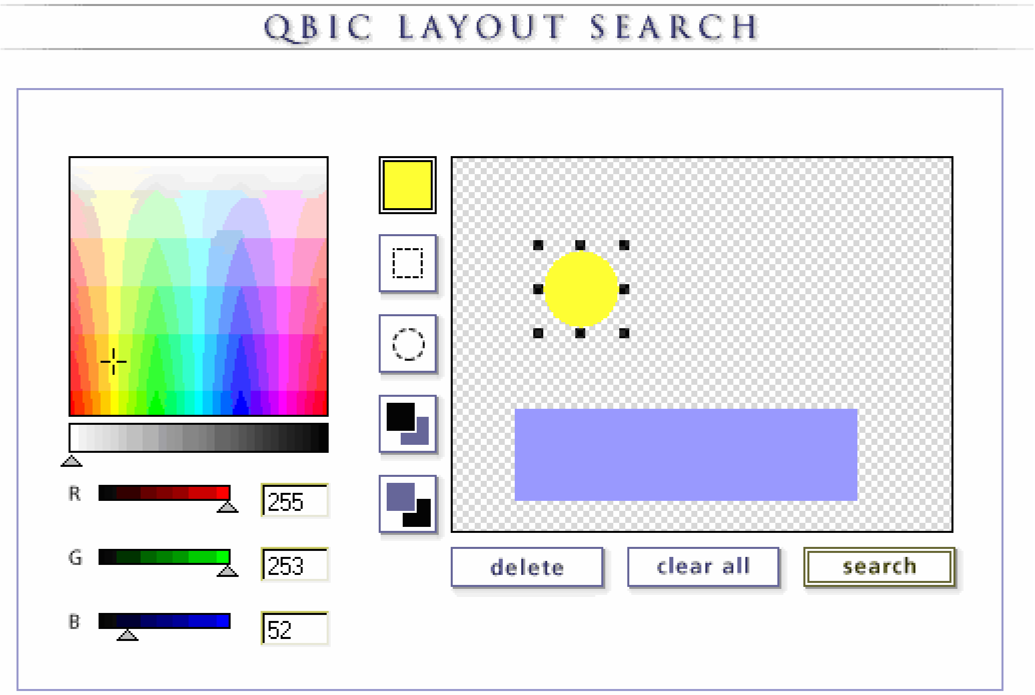


DATABASE TYPE: Photographs

MATCHING TYPE: Fuzzy

FEATURES: Color invariant

Figure 5. Example of using a sample image as a query in the PicToSeek CBIR system. (screen shot from http://zomax.wins.uva.nl:5345/ret_user/).



R 255

G 253

B 52

Figure 6. Example of the user interface in QBIC (screen shot from wwwqbic.almaden.ibm.com). This figure is better viewed in color. Users search images by selecting colors from a palette or by sketching shapes on a canvas.

1.3. Issues

Multimodal IR is a large research area. It is closely related to the following research areas: text retrieval, image retrieval, document image retrieval, and document clustering/classification. Each of these research areas is rather large and actively researched, so it is impossible to cover all the state-of-the-art. In this survey, we focus on the following issues:

- The basic indexing techniques and retrieval models for the text and image modalities in multimodal documents. The indexing techniques and retrieval models all aim to provide an effective and efficient IR.
- Adaptable and reconfigurable IR. Techniques in text retrieval have been adapted to image retrieval, or multimodal IR. State-of-the-art methods in other research areas may be adapted to multimodal IR. For multimodal IR, it is challenging to develop a general approach to combine text and images, which may be adapted to different types of documents or different modalities.
- Combination of text and images in multimodal IR. Retrieval using either the text or image modality has advantages and disadvantages. These modalities may be combined to overcome the limitations in a single-modality IR system.
- The relationship between collateral text and image. When describing image content, collateral text and image complement each other. There are also redundancies between text and images. One modality may help make the other modality more informative and more precise.
- Noise and uncertainties in IR systems. Uncertainties exist in various parts of IR systems. There are uncertainties when representing a user's information needs as a query, when indexing either collateral text or text from OCR, and when matching a query with documents. These uncertainties are dealt with in various ways.
- Techniques to improve effectiveness of IR, such as Latent Semantic Indexing, user's relevance feedback, semantic network, and document clustering and classification. These techniques are surveyed in Section 5.

2. Indexing techniques

Both textual content and visual content may be indexed in multimodal documents. Section 2.1 surveys textual content indexing, which usually uses a sequence of index terms to represent text. Various types of image features might be extracted to represent the visual content. Visual content indexing is surveyed in Section 2.2.

2.1. Textual content indexing

Text in an electronic document is assumed noise-free, in contrast with text in document images, where OCR has errors. Section 2.1.1 summarizes noise-free text indexing used in text retrieval. Indexing of collateral text used in text-based image retrieval is similar to noise-free text indexing. However, finding collateral text is not a trivial task in some documents. Techniques to find collateral text are reviewed in Section 2.1.2. Section 2.1.3 summarizes special techniques that are developed to index textual content in document images.

2.1.1. Noise-free text indexing

Index terms (or *keywords*) are the basic units to represent textual content of documents. In English natural language text, index terms are obtained using the following procedures [YaNe99]:

(1) Lexical analysis of the text is performed to extract candidate words to be used as index terms from the documents. During lexical analysis, various issues arise, such as how to deal with digits (*CISC 999*), punctuations (*depth.doc*), hyphens (*on-line*), etc.

(2) Optionally, stop words are removed. Stop words are words that occur too often to hold information, such as *the, of, in*. The removal of stop words usually improves precision, and reduces the size of the indexing structure considerably. It might reduce recall.

(3) Optionally, words are stemmed. Different word forms may bear similar meanings (e.g. *search, searching*). Stemming aims to group words with a common stem together. Stemming usually improves recall, and reduces the size of the indexing structure. In some cases it might reduce precision.

(4) Optionally, multi-word phrases are identified as additional indexing terms.

(5) Optionally, syntactic analysis of a sentence is performed and words from various grammatical classes are chosen, such as nouns, verbs. Usually nouns are chosen since they are assumed to capture most of the semantics of a document.

There is no conclusive evidence that these optional text processing steps yield consistent improvement in retrieval performance, especially when the collection of documents is general and heterogeneous [YaNe99]. That is the reason why some web search engines index all the words.

Indexing constructs an *inverted index* of word-to-document pointers. The inverted file is designed for rapid merging of document lists and calculation of similarities. We won't survey the data structures used for indexing.

2.1.2. Finding collateral text

Collateral text indexing is similar to noise-free text indexing. However, there are uncertainties when locating collateral text in some multimodal documents, e.g. web pages. We briefly address this issue since many multimodal IR systems in our survey retrieve images from web pages. Unlike journal articles, most images on a web page do not have an explicit caption. In web pages, the collateral text of an image may be extracted from many sources, such as the page title, the image file name, the alternate text and the surrounding text which include top, bottom, left and right directions [SwFA97].

The collateral text may be incomplete or irrelevant. If the surrounding text around an image is not considered (e.g. only titles in a news archive are used [ZhGr02]), the collateral text may be incomplete. On the other side, if using the full-text of the web page as the collateral text, some text may be irrelevant. Various techniques are developed to find surrounding text related to an image. For example, Mukherjea and Chost use criteria like visual and syntactic distance between images and potential captions to relate text to images [MuCh99].

Since the relevancy or importance of various text sources can be different, some systems give different weights to the text from different sources. In ImageRover, a word relevant to an image is identified based on its frequency of occurrence in the HTML text, its position with respect to the image, and its style [ScCS99]. Words with specific HTML tags are given higher weights. How to assign suitable relevance weights to different sources is a difficult problem.

2.1.3. Indexing textual content from document images

Due to the errors in OCR (Optical Character Recognition), robust indexing techniques are needed to index textual content from document images. Indexing of document images has been comprehensively surveyed in [Doer98] [MiCh00]. There are basically two indexing approaches:

- **Index textual content using OCR results**

  This type of indexing requires a full conversion of the text in document images using OCR. Many techniques have been proposed to deal with OCR errors. Lopresti uses an NGram method to index OCR text [Lopr96]. NGrams split up a document into n-character terms. For example, "DOCUMENT" is split up into several trigrams: DOC, OCU, CUM, UME, MEN, and ENT. Each trigram is an indexing term. This NGrams indexing produces a large size of indexing structure. Another approach is to correct OCR errors using a dictionary or language statistics, and then noise-free text indexing techniques are used [TaBC94]. This approach usually can't make a total correction of OCR errors.

- **Index images of text without OCR**

  OCR is error-prone for poor quality document imags (e.g. fax or documents from copier machine, historic documents, and handwriting). Textual content may be indexed directly from images without OCR. For example, Character Shape Codes based on character image features are used to index the textual content of the documents [SmSp97]. In the indexing procedure, first image segmentation is performed at word and character levels. Then each character in a word is mapped to a Character Shape Code without being identified. These codes together form a Word Shape Token for each word, which works as an index term. This approach can be cheaper than a full OCR conversion and may be more robust to noise. Also it works for all languages and scripts. It is unsuitable to deal with words with touching characters since characters can't be segmented. Alternatively, word image indexing may be used [LuZT04]. The image-level features of an entire word are coded without character segmentation and OCR.

2.2. Visual content indexing

The goal of visual content indexing is to extract characteristic image features and organize them in a way to facilitate CBIR. Multiple features may be extracted from an image and may be represented in multiple formats [RuHu99]. Image feature extraction borrows the research achievements from pattern recognition and computer vision. It is impossible to review all the state-of-the-art image feature extraction techniques. In Section 2.2.1, I summarize the most commonly used image features in CBIR. Some specific features, such as those used for medical images, human faces and finger prints, need domain-specific knowledge and they are covered elsewhere in the pattern recognition and computer vision literature. In Second 2.2.2, I survey special visual content indexing techniques for document images.

2.2.1. Visual content indexing in CBIR

Image features can be extracted at a global level, i.e. from a whole image, or a local level, i.e. from regions or objects of an image. To obtain the local image features, an image is often divided into parts first [RuHu99]. The simplest way is to partition an image into data independent parts. This kind of partition does not generate perceptually meaningful regions but is a way of representing the global features of the image at a finer division. An alternative method is to segment the image into salient regions according to some criterion, such as color and texture [CaTB99]. Region segmentation is not a trivial task. Unavoidably, there are uncertainties in region segmentation.

Color, texture, shape, and spatial layout features are the most commonly used visual features in CBIR. In the following, I provide an overview of these features.

Color

Color features are commonly used for retrieval of color images. They are easy to compute and are insensitive to small changes in viewing positions. Color features extracted from various color spaces (e.g. HSV, LUV) take various formats. Here, color histogram, color moments, and color correlogram will be discussed. A color histogram represents the distribution of the number of pixels for each quantized bin in each color channel. A normalized color histogram is used in ImageRover [ScTC97]. Stricker and Orengo use color moments to characterize the color distribution [StOr95]. Up to third order color moments, i.e. mean, variance and skewness, are extracted. A color correlogram is a three-dimensional histogram, of which the first and the second dimensions are the colors of any pixel pair and the third dimension is their spatial distance [HuKM97]. It is used to characterize not only the color distributions of pixels, but also the spatial correlation of pairs of colors.

Texture

Various texture features have been effectively used in CBIR systems combined with other features, such as in QBIC [Flic95] and in MARS [HuMR96]. The commonly used texture measures are: Haralick's gray level co-occurrence features, Tamura texture features and wavelet transform. Haralick's co-occurrence features represent the gray level spatial dependence of texture [HaSD73]. The co-occurrence matrix is based on pixel values at various orientation and distances. Statistics computed from the matrix include *contrast*, *inverse difference moment*, and *entropy*. The Tamura texture features include *coarseness*, *contrast*, *directionality, line-likeness, regularity,* and *roughness* [TaMY78].The Wavelet transform [Mall89] provides a multi-resolution approach to texture analysis and classification.

Shape

Shape features are extracted from an object or a region after image segmentation. Since robust and accurate image segmentation is difficult to obtain, shape features are not as commonly used as color and texture for a general collection of images. They have been limited to special applications where objects or regions are readily available, such as retrieval of line drawings [LoMo95] and retrieval of trademark images [JaVa98]. Mehtre et al compared various shape representations for CBIR [MeKL97]. In general, the shape representations can be divided into two categories: boundary-based and region-based. Boundary-based features use only the outer boundary of the shape. The most successful boundary-based features are Fourier descriptors. The main idea of a Fourier descriptor is to use the Fourier transformed boundary as the shape feature [PeFu77]. Region-based features use the interior of the shape region. The basic region-based features are centroid, area, eccentricity, circularity, and statistical moments.

Spatial layout

Spatial layout encodes the absolute or relative position of segmented objects or regions, for example, a blue region (e.g. sky) is on the top of a green region (e.g. grass). Spatial attributes can be represented in different ways, such as 2D-strings [ChSY87], spatial quad-tree [Same84], and symbolic images [GuRa95]. 2D-strings encode the spatial relationships of objects in horizontal and vertical directions using two strings. Symbolic images represent the original images logically by uniquely labeling image objects with symbolic names. Spatial relationships in a symbolic image are represented as edges in a weighted graph. Due to the difficulty of accurate segmentation, spatial features of segmented objects in an image are also used in limited applications.

Here we finish the overview of most commonly indexed image features in CBIR. These features are typically calculated off-line and stored for each image. There is no single optimal feature set for a large collection of heterogeneous images. Usually multiple types of features are extracted from the images. The critical problem and the open issue is how to dynamically choose a best set of robust features [RuHC97]. Goodrum points out that the early developments of CBIR systems focus primarily on the use of features that can be computationally acquired, but little has been done to identify the visual attributes suited for various tasks and collections [Good00]. For efficient retrieval in a large set of images, internal indexing structures for the feature data are also important. A survey of indexing structure is beyond the scope of this work.

2.2.2. Visual content indexing of document images

In document images, visual content may come from images within a document image. Indexing of images within a document is the same as visual content indexing in CBIR. A difference is that images in most CBIR systems are natural images in color (an example is Figure 5 in Section 1), so the color features are extensively researched. In contrast, most document images are either black-and-white or gray-scale images, so a different set of features are used.

A whole document image may be treated as a single image and indexed for retrieval based on visual similarity. Mostly-text document images have more distinctive structural characteristics than the natural or graphical images in CBIR. Figure 7 shows examples of mostly-text document images with distinctive structural features. Structural features are extracted from the results of document layout analysis. For example, a modified XY-tree representation is used for document page retrieval [CeMS02]. The XY-tree representation is a well-known approach for describing the physical layout of documents [NaSe84]. The root of an XY-tree is associated with the whole document image. The document is split into regions that are separated by white spaces. Horizontal and vertical cuts are alternately performed. Each tree node is associated with a document region.

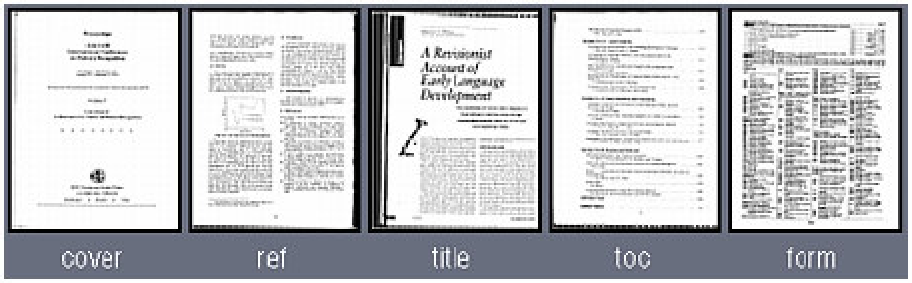


Figure 7.Examples of different types of document page images (cover, reference, title, table of contents and form) with distinctive structural features, reproduced from [ShDR01].

.

3. Retrieval models

A retrieval model describes how document representations and a query representation are compared to retrieve documents. In Section 3.1, we review text retrieval models. Text retrieval models are well-developed and they have been adapted to other branches of IR. In Section 3.2, we review CBIR retrieval models. We won't review the retrieval models in text-based image retrieval and document image retrieval since they are closely related to either text retrieval or CBIR retrieval. The text retrieval models may be enhanced to deal with noisy OCR text, such as approximate string matching. Multimodal retrieval models are a combination of text and CBIR retrieval models, and various combination techniques are developed, which will be surveyed in Section 4.

3.1. Text retrieval models

There are two major categories of text retrieval models: exact matching, e.g. Boolean model, and ranked retrieval, e.g. vector space model and probabilistic model. Exact matching matches the query with the documents. It assumes all matching documents are equally relevant to the query. It doesn't provide a scheme to rank documents in an order of estimated relevancy. Ranked retrieval models measure the degree of similarity between a query and a document, and return a set of documents ranked by how similar they are to the query. A vector space model ranks documents by geometric similarity to the query. A probabilistic model ranks documents by probabilistic relevance. They are not mutually exclusive. They provide similar solutions, but different interpretations for a same problem. Other models combine various retrieval methods. In the following, we introduce the vector space model and probabilistic model.

3.1.1. The vector space model

The vector space model [SaWY75] is the most popular text retrieval model. A collection of documents $D=(d_1, .., d_m)$ can be viewed as vectors in a very high dimensional vector space. Every term in the index language $(t_1, .., t_n)$ becomes an independent dimension (or a feature). A document d_i is represented as an n-dimensional vector $\overrightarrow{d_i} = (w_{1i},...,w_{ni})$. W_{ji} is the weight of t_j in $d_{i,}$ expressing how well feature t_j describes the content of document d_i. The query is regarded as a short document and is represented as a vector in the same space as the documents. The retrieval function computes the similarity between the description vectors of all documents with the description vector of a query, and returns the *k* documents ranked by the closeness of their vectors to the query.

Various similarity functions are used, such as Jacaard, dot product, and cosine similarity [YaNe99]. Cosine similarity is the most common one. For normalized vectors, the cosine is simply the dot product as shown in eq. (1).

$$sim(d_i, d_k) = \vec{d}_i \cdot \vec{d}_k = \sum\nolimits_{j=1}^{n} w_{ji} w_{jk},\ \vec{d}_i \textit{ and } \vec{d}_k \textit{ are normalized document vector} \qquad (1)$$

The vector space model returns ranked documents in an order that the documents most likely to be useful to the searcher are listed first. It benefits from queries with many terms, since not all of the terms in a query need to match the documents.

Term weighting

Proper weighting of terms affects the effectiveness of the vector space model. The index terms in a collection of documents have different frequency. There are three main factors when weighting terms: (1) *tf* (term frequency), the frequency of the term in a document. Terms that are frequent in a document are important indicators of the document's content. (2) *df* (document frequency), the number of documents containing the term. Terms that are frequent throughout the collection are not useful in distinguishing a document's content. Usually, *idf* (inverse document frequency) is used to represent the importance of a term. (3) Document length. Documents usually vary in lengths. A longer document tends to score higher because it uses more terms, and the same words may be used repeatedly.

Various methods for weighting terms have been developed. The most popular one is $tf \cdot idf$ weighting. $tf \cdot idf$ weighting has many variations. A simple $tf \cdot idf$ weighting is shown in eq. 2. The weights are usually normalized to discount the effects of document length.

$$tf \cdot \ln \frac{no.\ of\ documents}{df} \qquad (2)$$

3.1.2. Probabilistic models

Probabilities provide a principled foundation for uncertain reasoning. According to the *Probability Ranking Principle*, if documents are treated independently, the optimal ordering of returned documents is by decreasing probability of relevance [Rijs79]. The objective of a probabilistic model is to compute the probability that a document (D) is relevant given a query (*q*), i.e. *P(D is relevant| q)*. Various probabilistic models have been surveyed in [CrLR98]. They differ in how they evaluate the probability of relevance.

Language models for retrieval

Since 1998, probabilistic language models have become increasingly popular in text retrieval [PoCr98] [Hiem98] [MiLS99] [LaZh01]. Each document is represented as a generative probabilistic model of terms. Each document model defines a probability distribution over the terms in the vocabulary. The model is estimated from a sample of text representative of that model. Usually a unigram estimate of words is used, i.e. the terms are independent given a particular language model. Query terms are assumed to be independent observations from a document model. Documents are ranked by the probability *P(D|q),* the probability of a relevance of a document model given a query. It is the same as ranked by *P(q|D)* (the probability of a query generated by a document model), according to Bayes Theorem $p(D \mid q) = \dfrac{p(q \mid D) * p(D)}{p(q)}$, in which *p(q)* is constant, and the prior probability *p(D)* is often treated the same for all documents. Languages models must be smoothed so that non-zero probabilities can be assigned to query terms that do not appear in a given document. Various smoothing techniques have been used which affect the effectiveness of the language models. Berger and Lafferty adapt methods from statistical machine translation for the smoothing [BeLa99].

Language models are conceptually simple and self-explanatory. They have firm theoretical basis and the models make natural use of collection statistics, not heuristics. Advances already made in statistical natural language processing can be used. Although the term independence assumption is often not satisfied, the language models have been proved to be effective.

3.2. CBIR retrieval models

Given an example image or a sketch as a query, a CBIR system extracts the same visual features from the query as those used to index images. Similarity ranking methods are used to return a set of images ranked by their similarities to the query image. Both the vector space model and the probabilistic model in text retrieval are adapted to CBIR. Other special techniques are developed to retrieve images using various feature representations.

3.2.1. The vector space model

The vector space model is most commonly used in CBIR. The collection of images and queries are represented as feature vectors in an n-dimensional vector space. Unlike text retrieval, there is no single widely used similarity measure for CBIR. Antani et al. provide an overview of approaches for image similarity matching [AnKJ02]. Criteria for choosing a similarity measure include computational efficiency, and the ability to capture similarity between images. Different sets of features may use different similarity measures. For example, similarity of texture features is usually measured using Euclidean distance or Mahalanobis distance (which considers the correlation of different features). Histogram intersection (shown in eq. 3) has become popular for measuring color similarity based on color histograms [VaLi00]. The choice of similarity measure affects retrieval performance of a CBIR system significantly. Since a CBIR system may use different sets of features, choosing a suitable distance measure for each set of image features and optimally combining the similarities calculated from different set of image features become an important issue.

$$sim(h,g) = \frac{\sum_{j=1}^{n} \min(h[j], g[j])}{\min(\sum_{k=1}^{n} h[k], \sum_{i=1}^{n} g[i])} \text{ , h and g are color histograms with n bins.} \qquad (3)$$

3.2.2. Probabilistic models

Probabilistic models have been adapted to CBIR. An image is viewed as a generative probabilistic model that defines a probability distribution over visual features. When searching, the system measures the probability that a query image is generated by each model. A proper probability distribution is needed to capture the main characteristics of an image. Vasconcelos et al. built a separate Gaussian Mixture Model for each image [VaLi00] [Vasc00]. The image is cut into blocks of 8 x 8 pixels. 10 Discrete Cosine Transform (DCT) coefficients from each color channel are the features for each block. Each block is assumed to be generated by a mixture of Gaussian sources, where the number of Gaussian components is fixed for all images in the collection. An Expectation Maximization (EM) algorithm [DeLR77] is used to find parameters of the Gaussian Mixture Model. The number of Gaussian components is important. Too few components may not capture the most important aspects of an image. Too many components may cause over-fitting, i.e. there is not enough data to estimate the increasing number of parameters. Typically, a low number of Gaussian components (between 4 and 32) are used.

3.2.3. Other retrieval models

For features that are not represented in feature vector format, such as shape descriptors or descriptors of spatial layout, similarity measuring is more complex. It is related to object recognition in pattern recognition or computer vision. Veltkamp and Hagedoom provide an overview of shape matching techniques, such as matching based on turning angle functions, deformable template matching, and graph matching [VeHa01]. We won’t survey the details.

4. Combination of text and images

Text and images are distinct modalities used to represent contents in a document. They each have advantages and disadvantages, discussed in Section 4.1. While text and images may be separately ambiguous, jointly they are less so. This is because writers, when writing text, tend to leave out what is visually obvious in the image (e.g. the color of flowers) and mention properties that are very difficult to infer using vision (e.g. the species of the flower). Ingwersen's cognitive model of IR [Ingw96], which predicts that combining methods using different cognitive structures is likely to be more effective for retrieval than any single method, provides a theoretical basis for the combination of text and images.

While there is a substantial amount of completed and ongoing research in both text retrieval as well as image retrieval, much remains to be done to see how effectively these approaches can complement each other, and how the text and image modalities can be seamlessly integrated in a single framework. We discuss existing multimodal IR systems in Section 4.2 and summarize various text and images combination techniques in Section 4.3.

4.1. Advantages and disadvantages of text and image modalities

4.1.1. **Text modality**

Text retrieval has been successfully used in large and heterogeneous document collections, such as web page retrieval and digital libraries. The advantages are that the textual content from documents can be relatively easily represented by using a set of terms, and users represent their semantic information needs naturally using text.

Representing text by a set of terms unavoidably leads to some loss of semantics in the text. There are two fundamental problems when using text for retrieval:

- **Polysemy**: Words often have multiple meanings and different types of usage. For example, when a user uses a query "table" intending to find the furniture, the results may include irrelevant documents, where *table* is referred to as a data structure. This ambiguity problem becomes more severe in a heterogeneous collection of documents, such as web pages.
- **Synonymy**: IR may not retrieve relevant documents that include synonymous terms. For example, when a user wants to find information about cars, if the query contains only the term "car", relevant documents including "automobile" may not be retrieved.

Using text for image retrieval has the same advantages and disadvantages as those in text retrieval. Text-based image retrieval suffers more limitations than text retrieval. As we introduced in Section 2.1, there are uncertainties in finding collateral text blocks. Collateral text may be irrelevant and incomplete. Some images may not have collateral text at all. Some images are manually annotated with a few keywords. Manually annotating images is expensive and subjective. Also, annotations do not normally describe the visual content of an image, such as color, texture or shape. Sclaroff et al. indicate that some images could not be annotated because it is difficult to describe their visual content with words [ScCS99].

4.1.2. Image modality

An image is worth a thousand words. On the one hand, this saying indicates that images are a powerful modality for communicating information; on the other hand, it says images are complex to describe.

An image can be described in various ways, such as its visual content (i.e. what it looks like), or its semantic content (i.e. what it is about). Visual content indexing is surveyed in Section 2.2. When searching images, different users may have different information needs for the same image. For example, for a bar chart showing the statistics of the population, it might meet the information needs of a user looking for bar charts based on visual content, or the needs of another user looking for information on the population statistics based on semantic content.

CBIR allows users to search for images based on image visual content, such as color, texture, and shape. Strengths of CBIR systems are that the indexed features can be domain-independent, i.e. the features of color, shape, and texture are not limited to a particular image domain. In addition, the automatic indexing provided by these systems is efficient without human subjectivity.

CBIR has limitations. It has not been used as widely as text retrieval. One of the fundamental problems is the *semantic gap,* i.e. lack of a link from semantic information needs to visual content [SmWS00]. Rretrieval results based on the similarities of pure visual content do not necessarily possess semantic meanings that are of interest to the user. For example, a query using an example image of a green apple based on color features might return images of grass, an apple or a green-coloured wall. The processes of grouping image features into meaningful objects and attaching semantic descriptions to them are difficult. In current practice, the semantic contents of images are obtained from collateral text.

The *page-zero* problem exists in CBIR [ScCS99]. If the user doesn't have an example image available, they find it is hard to start the retrieval. In some systems, a complex query interface has to be designed to let the user draw a sketch, e.g. the query interface in Figure 6 (Section 1). A complex query interface burdens the users.

Multimodal IR may combine the strengths from both the text and image modalities, and overcome their limitations. Use of images can help tackle ambiguity in text retrieval. Use of text can help narrow the semantic gap in image retrieval.

4.2. Overview of multimodal IR systems

Multimodal IR systems differ in the complexity of the document models being used, and the combination methods. Table 1 provides an overview of multimodal IR systems that have been surveyed. The first column shows the reference and the experimental system's name if available. We cluster the surveyed multimodal IR systems using the following criteria:

- Complexity of the document models (second column in the table)

  Most document models in the surveyed multimodal IR systems are simple, in which a document is a single image annotated with keywords. There are a few complex multimodal document models [OrPM99] [MeSS01]. Those models are similar to the document model presented in Figure 2 Section 1. The documents contain both text and images, and the text includes independent text and collateral text.

- Text and image combination techniques (third column)

  Text and image combination techniques are surveyed in Section 4.3. In general, there are three approaches:

  (1) The text and image modalities are sequentially used.

(2) The text and image modalities are simultaneously used, either linearly or nonlinearly combined.

(3) Probabilistic methods are used to integrate text and images.

- Whether relevance feedback is used and how it is used (fourth column)

User's relevance feedback is an important technique to improve effectiveness of IR. It is surveyed in Section 5.3. Some systems rely on user's relevance feedback to combine text and images, while some systems don't posses relevance feedback function.

- Image features used in the systems (fifth column)

Various image features are indexed in the surveyed systems as described in Table 1. Textual features are similar in all the systems; they use a set of keywords, and most systems use $tf \cdot idf$ weighting (discussed in Section 3.1.1). The main difference between systems is the degree of Natural Language Processing that is used. For example, in some systems, text blocks are further processed with Part-of-Speech tagging to get nouns. We won't provide the details of textual indexing in multimodal documents.

Table 1. An overview of multimodal IR systems.

	Document model	**Text and image combination**	**Relevance feedback**	**Image features**
[BaFo01]	Images with keywords.	Probabilistic model (hierarchical).	No.	Features from segmented regions (color, orientation energy, region size, location, convexity, first moment).
[BlJo03]	Images with keywords	Probabilistic model (LDA).	No.	Features from segmented regions (size, position, color, texture and shape).
[BrCJ97] DocBrowser	Faxed business letters.	Sequentially.	No.	Various visual features from logos, signatures, etc.
[ChLZ01] *iFind*	Web images with collateral text.	Linear combination.	Yes. Refine query and learn a semantic network.	Global features. Color histogram; color moments, color coherence, wavelet texture, Tamura texture.
[JeLM03]	Images with keywords.	Probabilistic model (cross-media relevance model).	No.	Features from segmented regions. Same as those used in [DuBF02]. 33 region features.
[MeSS01]	Structured documents containing text and images.	Non-linear combination. Fuzzy logic reasoning.	No.	Shape or color.
[NaKT03]	Image with keywords.	Linear combination.	Yes. Re-weighting.	Color histogram, texture features such as regularity, coarseness, and orientation histograms.
[OrPM99] WEBMARS	HTML documents.	Linear combination.	Yes. Query refinement.	Multiple sets of features including color, texture and layout features.

[ScCS99] ImageRover	Web images with collateral text.	Linear combination with latent semantic indexing (LSI) applied to text.	Yes. Re-weighting.	Color histogram and the dominant orientation histogram.
[ChSB97] VisualSeek WebSeek	Web images with collateral text.	Sequentially.	Yes. Query refinement.	Color histogram.
[Srih00] MMIR	Web images with collateral text.	Linear combination.	No. Off-line learning.	Face detection techniques, color histogram.
[WaMX04]	Web images with collateral text.	Non-linear combination. Iterative similarity propagation.	No.	Color correlogram, color moment, wavelet textures.
[ZhGr02]	Web images with collateral text (only web page titles).	Latent semantic indexing applied to unified image and text feature vector.	No.	Color histogram in HSV color space and color anglograms.
[ZhHu02]	Images with keywords.	Linear combination	Yes. Word Association via Relevance Feedback (WARF).	Color moments, wavelet moments and edge-based structure features.

4.3. Combination techniques

Most of the research on combining text and images is performed in the image retrieval research community, where collateral text is combined with images to improve the effectiveness of image retrieval. The theme of the research in the surveyed papers is to discover the relationships between the collateral text and images. Based on the relationships between the text and image modalities, the collateral text may by used to find the semantic concepts of the image and propagate semantic concepts to unlabelled images using both the text and image modalities. Text and images may be associated implicitly, e.g. using Latent Semantic Indexing (surveyed in Section 5.1). The relationships between text and images may be discovered by creating semantic networks (surveyed in Section 5.2).

In the text retrieval community, little research is done to use images to improve text retrieval. As shown in Figure 2 of Section 1, current text retrieval uses only text blocks and ignores images. If a relevant image exists in a multimodal document, it might contribute a lot to the final ranking of the document. Paek and Smith investigate some issues related to this question [PaSm98]. One issue is judging whether an image is related to the textual content of the page. Images are classified into content images (related to page text) and non-content images (such as advertisement and logo images). The content images are used in the summary of the document.

In this section, we survey various techniques to combine collateral text and images.

4.3.1. Sequential use of the text and image modalities

The text and image modalities may be used sequentially and only one modality is used in each query. In essence, the text and image modalities are not integrated in this approach. They are indexed and retrieved separately. Many IR systems, including web page retrieval systems or document image retrieval systems that allow the users to search either modality use this kind of approach. Either text or image modality may be used first, depending on the role of the two modalities.

Text based image retrieval may be used first to provide an initial set of images to start CBIR. CBIR is used to refine the results of the text based image retrieval. The text modality may be used to categorize the documents into semantic categories. In WebSeek, collateral text is used to perform semi-automatic classification of images into a taxonomy of semantic categories [ChSB97]. The user can browse a set of images within a category. Color histogram based similarity matching is used to find images with similar color features within a category or over the entire catalog. CBIR may be performed first, and then the text modality is used to refine the query results or provide a better browsing of the query results [MuCh99].

Retrieval of both text and images are provided in some document image retrieval systems. For example, DocBrowser retrieves faxed business letters [BrCJ97]. Visual content includes the company logo on a letter, handwritten signatures, entire document pages, and words not identified by the OCR engine. Textual content is the OCR results. Both text content and line-based features are also used to retrieve line drawings in [LoMo95].

In special cases, the text modality may help image indexing. In the PICTION system [SrCB94], the caption of an image is first analyzed to identify the expected number of faces and their expected relative positions. Then a face detector is applied to a restricted part of the image. Similarly, in the MARIE project, captions are analyzed using natural language analysis techniques to help identify shapes in the images [Rowe95] [Rowe99].

4.3.2. Both modalities are used simultaneously

Both the text and image modalities may be used simultaneously when performing a query. A composite query is formed to include both text and an image. If no example image is provided, some systems use a random vector for visual feature vectors in order to let a user explore different parts of the feature space [ZhHu02]. During retrieval, the system combines the similarities computed from the text and images. The text and images may be indexed separately. There are various combination methods. They are grouped into two general categories: linear combination and non-linear combination.

4.3.2.1. Linear combination

Linear combination is a simple method and is most commonly used. A composite query q has two parts: text query q_T and image query q_M . The similarity between a query q and a multimodal document D_i , which also has two parts, text D_{i_T} and images D_{i_M} , is defined in the following:

$$Sim(q, D_i) = \alpha Sim_T(q_T, D_{i_T}) + (1-\alpha) Sim_M(q_M, D_{i_M})$$

The combined similarity is the weighted sum of similarities computed from both modalities. The weight α is the *inter-modality weight*. *Intra-modality weights* are weights within the text or image feature space. The effectiveness of linear combination depends on the choice of weights and other issues. Different systems differ in the computation of similarity measures for text and images, and the weight α to combine the text and images.

There are three general ways to set the *inter-modality weight*. (1) The weight may be set manually according to prior knowledge. For example, the two modalities are given equal weight to linearly combine similarities from textual features (using dot products) and visual features (using Euclidean distances) in the *iFind* system [ChLZ01]. (2) The weight may be learned off-line. For example, to find the optimal weight set for the multiple modalities, a training phase is used to learn a group of optimal weights for a selected set of representative queries in the MMIR system [Srih00]. In the retrieval phase, the individual models for different modalities are linearly combined using the weight set of the representative query which is the most similar to the current user-submitted query. (3) The weight may be adjusted on-line based on user's feedback [OrPM99] [ZhHu02]. User's relevance feedback is further surveyed in Section 5.3. The methods to set the *intra-modality weights* are the same as the weight adjusting methods used in either text retrieval or CBIR.

4.3.2.2. Non-linear combination

Non-linear combination is used in some systems. In the following, we summarize some representative non-linear combination techniques.

- **Fuzzy logic**

  Some systems use fuzzy logic reasoning to combine multiple modalities [MeSS01] [Sant02]. Meghini et al. combine features pertaining to text, images and document structures [MeSS01]. The document contents and queries are described using a logic language, Description Logic [Borg95]. The authors claim that the Description Logic languages have an object-oriented characteristic that makes them especially suitable for reasoning about hierarchies of structured objects. Hamacher sum, a disjunction operator ($\vee$) in fuzzy logic is used in [Sant02].

- **Iterative reinforcement**

  An iterative approach is used to explore the mutual reinforcement between images and their collateral text [WaMX04]. First, a similarity matrix of text blocks and a similarity matrix of images are created from the document-term matrix and the image-feature matrix respectively. Iterative reinforcement is performed based on eq. (5), adapted from the approach proposed by Kandola et al. [KaTC02]. The similarity in one modality is propagated to the other modality. Reinforcement can enhance or reduce the similarity of two objects. Both single-modality retrieval and linear combination of multiple modalities are special cases of eq. (5). If α =1 or β=1, it is a single-modality retrieval. If λ=1, it is a linear combination. When λ is set less than 1, the propagated similarity is weaker than the original similarity. In their experiments [WaMX04], the iterative reinforcement is proved to outperform linear combination, and to perform better than the single modality retrieval using either text or visual features.

$$\begin{cases} \hat{K} = \alpha K + (1-\alpha)\lambda Z \hat{G} Z^T \\ \hat{G} = \beta G + (1-\beta)\lambda Z^T \hat{K} Z \end{cases} \tag{5}$$

K (MxM) is the similarity matrix of M images.
G (NxN) is the similarity matrix of N text blocks.
Z (MxN) is the link matrix. A link exists if an image is associated with a text block.

4.3.3. Integration in a probabilistic model

Text and images may be seamlessly integrated in a probabilistic model. The reason we put this topic in a separate section is that the probabilistic models not only combine the retrieval results from the text and image modalities simultaneously, but also associate text with images. The association of text with images using another approach, semantic networks, is discussed in Section 5.2. The probabilistic models are adapted from those in single-modality IR systems introduced in Section 3. The probabilistic models have many variations. They may differ on the assumption of prior probability, and on the methods to estimate the probability distributions from both text and images. Some models use image features from segmented regions, while some models use image features without performing image segmentation. In the following, we summarize some representative probabilistic multimodal IR models.

The probabilistic models estimate the joint distribution of text and images P(T,M). Most of the models assume the text and image modalities are dependent. From the joint distribution, the conditional distribution of images given text P(M|T) and text given images P(T|M) can be obtained. Thus the models can be used for cross-modal information retrieval, for retrieval with a composite query, and for automatic image annotation. For example, Jeon et al. develop a cross-media relevance model (CMRM) [JeLM03]. They adapt techniques from cross-language retrieval [LaCC02]. Every image is described using a small vocabulary of blobs, which are clusters of all the segmented image regions in a collection of images. The blob generation algorithm is the same as the one used in [DuBF02]. Since an image has a similar representation as text data (an image is a set of blob numbers), language models are used to describe both textual and blob distributions.

Some probabilistic models are more advanced. For example, a Latent Dirichlet allocation (LDA) model is proposed in [BlJo03]. It assumes that a Dirichlet distribution can be used to generate a mixture of latent factors (clusters), and this mixture of latent factors is then used to generate words and regions. Based on the LDA model, the conditional probability of words given an image *p(T|M)* can be computed. For an N-word query *q*={q1,..qN}, each image M computes its score relative to the query as $\prod_{n=1}^{N} p(q_n \mid M)$. The model can be used for cross-modal information retrieval. It can retrieve un-annotated images using a text query. A probabilistic retrieval model based on a hierarchical clustering of images and words is proposed by [BaFo01]. The model learns the joint distribution of image regions and words. The documents belonging to a given cluster are modeled as being generated by the nodes along the path from the leaf corresponding to the cluster, up to the root node, with each node being weighted on a document and cluster basis. A query is the union of query words and query blobs.

5. Techniques to improve effectiveness of IR

Various techniques have been developed to improve effectiveness and efficiency of IR systems. In this section, we survey several representative techniques used in multimodal IR, including Latent Semantic Indexing, semantic networks, user's relevance feedback, and document classification/clustering.

5.1. Latent Semantic Indexing (LSI)

5.1.1. Introduction to Latent Semantic Indexing

Latent Semantic Indexing (LSI) is originally used to tackle the synonymy and polysemy problems in text retrieval; it relates documents that use different terms [DeDF90]. The general idea is to map documents and terms to a low-dimensional representation. The mapping is done by applying Singular Value Decomposition (SVD) to the document-term matrix, and selecting only the first a few (typically 100-300) eigen-vectors with the highest eigen-values. This ensures that the low-dimensional space (called *latent semantic space* or *concept space*) reflects semantic associations, since similar terms are mapped to similar locations in the latent semantic space. A query *q* is also mapped into this space. Document similarity between a query and documents is computed based on the inner product in this latent semantic space and it is invariant to the choice of words for similar concepts.

LSI tries to retrieve documents based on the concepts of the documents, instead of matching directly on keywords as done in classical text retrieval models, surveyed in Section 2.1. LSI usually improves recall, and it also improves precision. It has limitations. The axes in latent semantic space are not understandable by humans. The SVD algorithm is expensive, with complexity $O(N^2k^3)$. N is the number of terms plus documents and k is the number of dimensions in latent semantic space. Determining the optimal number of dimensions (k) is not trivial.

5.1.2. LSI in multimodal IR

Latent Semantic Indexing has been adapted to multimodal IR. There are two approaches: LSI only applied to text, and LSI applied to the unified feature vector containing both text and image features.

5.1.2.1. LSI applied to text

In this approach, LSI is first applied to text. Then the similarity computed based on the text feature vectors is linearly combined with the similarity computed from the image feature vectors. This approach is used in ImageRover to capture image-relevant text statistics from HTML documents [ScCS99]. Reported experiments show that maximum retrieval performance is achieved when both visual and textual content are employed. Sometimes, the performance is poor when there is a lack of correlation between the image content and the surrounding text (such as banners and logos in a web page). Since LSI is used only on text, this approach can't capture the co-occurrence between the image features and text keywords.

5.1.2.2. LSI applied to both text and images

In this approach, first text and image feature vectors are combined, then LSI is applied to the unified feature vector [ZhGr02] [West00]. Image feature vectors are treated differently in different systems. Real-valued image feature vectors from color histogram statistics may be directly used directly [ZhGr02]. The dimension of the image feature vectors is 100. Since only a small data set is used, the dimension of the text feature vectors is only 43. This is unusual; typically, the dimension of text space is a few thousand. Color and texture features may be quantized [West00]. The purpose of the quantization is to obtain a discrete set of tens of thousands of visual ‘terms’ that can be either present (one or more times) or absent in a document. A quantization technique proposed by Squire et al. may be used to ensure that the obtained image feature terms have a similar type and distribution as text terms [SqMP00].

Applying LSI to the unified feature vectors can implicitly associate some image features with the set of co-occurring keywords. However, it has limitations. It is difficult to combine multiple sets of feature into a single feature vector since as we discuss in Section 2.2, different types of features may require different similarity metrics.

5.2. Semantic networks

5.2.1. Introduction to semantic networks

Semantic networks were first used in text retrieval to ameliorate the synonymy problem discussed in section 4.1.1. A semantic network shows the terms and their relationships. A query may be expanded using words related to the query terms. The semantic network may be a general purpose thesaurus, such as WordNet [Mill95]. However, it has been shown that query expansion using automatically selected synonyms from a pre-built thesaurus yields poor results on a large, heterogeneous collection of documents [Voor94]. This is because relationships captured in such a thesaurus may not be valid in the local context of a given user query. A semantic network may be created via user’s relevance feedback. A statistical analysis on term occurrence and co-occurrence in the set of relevant documents is performed to automatically construct a thesaurus [Salt89].

5.2.2. Semantic networks using both the text and image modalities

In multimodal IR, semantic networks can be created using both the text and image modalities. Based on the assumption that semantically related images may be visually similar and vice versa, the semantic networks in multimodal IR link keywords with images or regions of images. The relationships among keywords are created by using a combination of text and image features.

Semantic networks are believed to play more important roles in multimodal IR than in text retrieval. Semantic networks may be used to expand a query in various ways: expand a text query using related keywords, expand a text query using visual content related to the keywords, and expand an example-based image query using text related to the images. In addition to query expansion, the semantic networks have other applications. They may help disambiguate the collateral text since the collateral text may be incomplete, irrelevant or subjective. The semantic networks may help narrow the semantic gap. They may also help in cross-modal retrieval, i.e. using text to retrieve images without annotations.

We distinguish two types of semantic networks, flat and hierarchical. Flat semantic networks assume a simple relationship between images and keywords, i.e. all the keywords are directly associated with the images or visual features of the images. No hierarchical or other relationships exist among semantic keywords. In hierarchical semantic networks, the keywords in a higher level are not directly linked with images, but via lower-level keywords. The early efforts of building a semantic network relied on a lot of user interactions and manual work. For example, in Chabot, the users can define concepts via an interface [OgSt95]. The concepts are associated with a combination of visual and textual predicates that define their visual representation (e.g. a *yellow flower* concept is associated with the *flower* keyword and the *yellow* color). To create a semantic-visual thesaurus, the users may need to manually label the image regions with visible keywords [Duff98]. In the following, we survey techniques that automatically create semantic networks.

5.2.2.1. Flat semantic networks

A flat semantic network proposed by Lu et al. [LuZH03] is shown in Figure 8. Many keywords may be associated with an image or regions in the image with different weights. The weights represent the semantic relevance of a keyword to an image. The links between the keywords and the images may be created by propagating a few manually annotated images to the other unlabelled images (called *automatic annotation*) [ZhSu03], or by using the relationship between the collateral text and images [LuZH03].

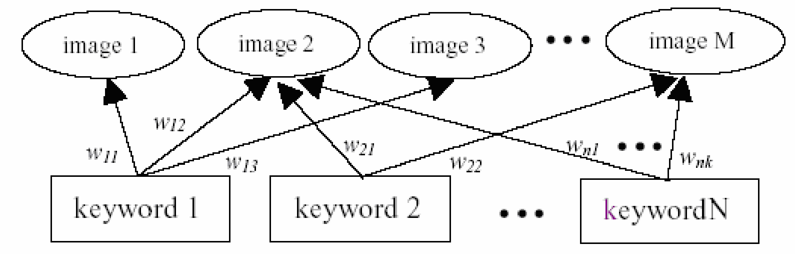


Figure 8. A flat semantic network [LuZH03].

Different techniques have been used to adjust the weights from the keywords to the images. Lu et al. use a simple voting scheme to update the weights of each link based on user's relevance feedback [LuZH03]. The basic idea is to assume the query terms used by a user represent the semantic meanings of the relevant images. The links from the query terms and the relevant images are added into the semantic network. If the terms are already in the semantic network, then their weights are increased. The weights of the terms used for the negative images are decreased. If the weights are less than a threshold, the terms may be removed from the semantic network. Using relevance feedback, a semantic network may be adaptive to an individual user's profile. In order for the voting procedure to be effective, a significant number of user queries and feedback iterations are necessary. Probabilistic methods are also used to modify the weights of keywords [ZhSu03] [BrZi04].

Semantic networks may be created based on the probabilistic models surveyed in Section 4.3.3. From the joint probability of text and image, the conditional probability of text given an image can be obtained; this is used to calculate the weight from a keyword to an image or a region.

5.2.2.2. Hierarchical semantic networks

Hierarchical semantic networks distinguish keywords associated with images at different levels [ZhZO04]. One level is *visible keywords* (also called *perceptual concepts),* such as *cloud.* This level has direct connections with image visual features. The higher level is abstract keywords (called *semantic concepts),* such as *vocation*, *holiday*. These words don't have direction connections with image visual features. The abstract words are realized through *visible keywords*. The different degrees of complexity regarding a concept are analyzed in [GoCL04]. Some concepts are diverse, e.g. the flowers concept, which encompasses flowers of different colors and types. Some concepts are well isolated, e.g. Eiffel Tower.

Various methods are used to build hierarchical semantic networks. The networks are created either based on global analysis of the whole document collection or based on relevance feedback.

- **Semantic network via global analysis**

In this approach, hierarchical semantic networks are created based on the analysis of a whole collection of multimodal documents. An example is the MeidalNet [BeSC00] [BeCh02a], which combines perceptual and semantic concepts in the same network. Image features are extracted from the whole image and the segmented regions. Perceptual concepts are discovered by clustering images based on their visual and text features. Semantic concepts are extracted by disambiguating the senses of words in annotations using the lexical database WordNet and image clusters. In addition, the relationships between keywords can be extracted using relations established in WordNet. Relationships among perceptual and semantic concepts are found based on co-occurrence. Images in the same cluster are assumed to be semantically related. Using the semantic network, low level feature queries can be translated to high-level semantic queries and vice versa. Nakagawa et al. create a similar semantic network [NaKT03]. They propose a novel multi-scale segmentation framework to detect prominent image objects. Segmented image objects are clustered according to their visual features and the created clusters are mapped into related words determined by psychological studies. A hierarchy of words expressing higher-level meaning (such as female, person, living thing) is linked to the lower-level words. They also rely on WordNet to find the conceptual words.

Using the general purpose thesaurus WordNet in multimodal semantic network suffers the similar limitation as we discussed in Section 5.2.1. The semantic associations are context dependent. Different users at different times may have different interpretations or intended usages for the same image, which makes fully automated off-line preprocessing (e.g. clustering, or classification) impractical in general. To ameliorate these limitations, the weights in the semantic network may be updated using relevance feedback [NaKT03].

- **Semantic network via relevance feedback**

One of the methods to build hierarchical semantic network is Word Association via Relevance Feedback (WARF) [ZhHu02]. An initial semantic network is created based on the image annotations. Visually similar images are assumed semantically related. For example, given two positive car images with different annotations, *Ford* and *car*, this method will associate *car* with *Ford*. After a user's feedback, the relevant relationship of terms *i* and *j*, *Sij* is updated as:

$$S_{ij} = S_{ij} + \max(f_i, f_j) \times (\min(f_i, f_j) - c_{ij})$$

f_i is called the *relevant term frequency*. It is the number of occurrences of a relevant term *i* in the relevant set. c_{ij} is the number of co-occurrences of two relevant terms *i* and *j* in the same image. The formula assumes that if two terms appear in the annotations for the same image, no association information for these two terms can be obtained out of this fact. Multiplication is used in the formula based on the assumption that the two terms are more likely to be relevant when they appear in a similar number of relevant images. The word similarity matrix may be specific to the data sets and the users. This similarity matrix is used for keyword semantic grouping using Hopfield network or clique detection, automatic thesaurus construction and soft query expansion. The relevance between a keyword *i* and an image *j* is computed as follows:

$$w_{ij} = \max S_{ik}\ , k \in \{k \mid keywords\ used\ to\ annotate\ image\ j\}.$$

This says that the relevance of a keyword to an image is equal to its association with the most relevant term used to annotate the image.

5.3. User's relevance feedback

5.3.1. Introduction to relevance feedback

Relevance feedback is a well-known technique for improving effectiveness of text retrieval [SaBu90]. It releases the burden from the user to form an effective query in the beginning. The system formulates a better query based on a user's feedback after an initial query. Relevance feedback attracts more research efforts in CBIR than in text retrieval. One reason is that it is easier for the user to judge whether the retrieved images are relevant than the text documents since images reveal their contents to users instantly. Another reason is that the *semantic gap* in CBIR makes relevance feedback more necessary than in text retrieval. Overviews of interaction techniques in image retrieval are given in [WoSS00] [ZhHu03].

The main idea of relevance feedback in text retrieval is to refine the original query by adding new terms from relevant documents (i.e. query expansion) and enhance the importance of query terms appearing in relevant documents (i.e. term re-weighting). The query refinement is done on a per-user basis. One of the best known query refinement approaches is the Rocchio algorithm [Rocc71]. The modified query vector is shown in eq. (6), where $\vec{q}_m$ is the modified query vector; $\vec{q}_0$ is the original query vector; α, β and γ are weights; D_r is a set of known relevant document vectors; D_{nr} is a set of known irrelevant document vectors.

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j \qquad (6)$$

This algorithm aims to move a new query toward relevant documents and away from irrelevant documents. The Rocchio algorithm has been adapted to CBIR and multimodal IR.

Relevance feedback is also viewed as a binary classification problem [ZhHu03]. The relevant documents are labeled as positive and the irrelevant documents are labeled as negative. The goal is to learn a classifier to better classify the documents to be searched. Many issues exist in such a classification task, such as small training samples and high-dimensional feature space. These issues are actively researched in CBIR research community.

5.3.2. Relevance feedback in multimodal retrieval

The use of relevance feedback in multimodal IR is similar to its use in text retrieval or image retrieval. The difference is that both text and image modalities are used. For example, Lu et al. modify the Rocchio algorithm to incorporate both text and image content [LuZH03]. Chen et al. use the relevance feedback to adjust the *intra-modality and inter-modality weights* (introduced in Section 4.3.2.1) [ChLZ01]. They performed experiments based on the images collected from different web sites with various contents, and they observe that content-based image retrieval based on low-level features yields relatively good results in some categories such as *sun* and *waterfall*. But in other categories such as *Clinton* and *summer*, the results are not so good since these words are more abstract. Different weights are given to text or image modalities with regard to different queries. The weights are also adjusted among words in the text features. ImageRover uses the user's feedback to select the appropriate similarity metrics for different types of image features, and adjust weights to different modalities [ScCS99].

As surveyed in Section 5.2, relevance feedback is also used to refine semantic networks.

5.4. Document clustering and classification

5.4.1. Introduction to document classification and clustering

Document clustering and classification are related to IR. They are useful for organizing and browsing documents in IR systems. Document clustering and classification share many characteristics with IR [Seba02]. We believe the techniques used in multimodal IR and multimodal document clustering/classification are much related and they will contribute to each other.

Document classification (or supervised categorization) is supervised learning, in which a set of classes is defined and training documents for each class are labeled. Various classifiers exist, such as KNN (K-Nearest Neighbor), naïve Bayes, and Support Vector Machine. Sebastiani provides a comprehensive survey of text categorization [Seba02]. In previously published work, we survey document image classification based on image features, document layout features, textual features and combination of various features [ChBl04]. We summarize the applications of document image classification, and identify important issues in designing a document classifier, including the definition of document classes, the choice of document features and feature representation, and the choice of classification algorithm and learning mechanism. Image classification is used to assign semantic concepts to images. We view image annotation as a task of image classification. The difference is that multiple labels may be assigned to an image or regions of the image [MiPi95] [WaLW01].

Document clustering is unsupervised learning, i.e. no labeled training samples are required. It groups documents into sets of similar objects. Numerous document clustering algorithms appear in the literature [Will88] [JaMF99]. These algorithms can be classified into two groups – those producing hierarchical clusters, such as Hierarchical Agglomerative Clustering (HAC), and those producing a flat partition, such as K-means clustering, and Self-Organizing Map (SOM) algorithm.

Document clustering and classification are large research areas. They have been used extensively in single-modality documents, either text or images. We won't survey single-modality document clustering or classification.

5.4.2. Document classification or clustering combing text and images

We will give a brief overview of document clustering or classification techniques that combine text and images. Similar to multimodal IR, the results of clustering or classification combining text and images are claimed better than the results using either text or image features alone [PaSH99] [BaFo01]. There is not much research on combination of text and images for classification or clustering as we survey so far.

5.4.2.1. Multimodal classification

It is challenging to automatically assign semantic labels to images. Efforts have been made to use both text and images to classify images [PaSH99] [SwFA97] [GeAl03]. In the current image classification systems, only a few semantic classes are usually defined. For example, the system proposed by Paek et al. separates images into two main semantic concepts, indoor and outdoor [PaSH99]. The text features are terms extracted from image captions. Weighted summation is used to combine the scores computed from the text and image features. WebSeer classifies images into photographs, portraits and computer-generated drawings using both image content and collateral text from web pages [SwFA97]. A distinct exception is the classification system in WebSeek [ChSB97]. A customized semantic ontology is defined for a general collection of web images (about 650,000 images) and more than 2,000 semantic classes are defined semi-automatically.

Text and image modalities may be applied separately instead of simultaneously in a classification system. For example, a hierarchical classification scheme is proposed by Lu and Drew [LuDr01], in which text and image features are used sequentially at different levels of classification.

5.4.2.2. Multimodal clustering

The text and image modalities are either used sequentially or simultaneously in the multimodal document clustering systems that we survey. An example of a sequential use of multiple modalities is the Scatter/Gather system [ChGN99]. The system helps a user progressively narrow a collection to a small number of elements of interest, similar to the Scatter/Gather paradigm for text documents browsing [CuKP92]. Scatter/Gather iteratively refines a search by "scattering" a collection into a small number of clusters, and then a user "gathers" clusters of interest for "scattering" again. A set of features, from different modalities, is pre-computed for each document image and stored as vectors. The text features include the words of text surrounding and associated with each image, the URL of the image, alt tags, and hyperlink text. The image features include a color histogram and a measure of color complexity. The documents are initially clustered into groups based on the text features. Clustering is performed using a standard k-means clustering algorithm with a preset number of clusters. Then the clusters selected by the user are re-clustered based on image features.

In some systems, the text and image modalities are used at the same time. For example, both text and low-level image features are integrated into a feature vector to cluster images in [BeCh02b]. Each cluster is considered a perceptual concept. They show that both visual and text feature descriptors are useful in extracting perceptual knowledge from annotated images. A hierarchical clustering system is proposed in [BaFo01]. It is extended from Hofmann's Hierarchical Aspect Model for text [Hofm98] and uses both text and image features. In the system, images with the associated word *flowers* may be broken into associated clusters which have predominantly red flowers, predominantly yellow flowers, and predominantly green garden scenes.

6. Conclusion

In this survey, we focus on multimodal document retrieval combining the text and image modalities. Multimodal IR is based on single-modality IR systems, including text retrieval, CBIR, text-based image retrieval, and document image retrieval. We survey the state-of-the-art in indexing techniques and retrieval models of various single-modality IR systems. Retrieval based on text modality is well researched and techniques in text retrieval have been adapted to other IR systems. Retrieval based on image modality is more complex than retrieval based on text modality. An image can be described at various levels using different sets of features. A semantic gap exists between user's semantic information needs and visual content of images.

A diagram of a multimodal IR system is shown in Figure 9, which summarizes various issues surveyed. A Multimodal IR system indexes both textual and visual content. The query space represents a set of possible queries. The query space might be Boolean combinations of words and phrases, as in the query "*sunset* AND *birds*"; or the query space might be images, as in "find me an image that looks similar to this one"; or the queries may be formulated as composite queries, containing both text and images. Indexing depends on the type of queries that are expected. The query formulation also depends on the types of queries. When retrieving, various techniques are used to combine results from the text and image modalities. Multimodal IR aims to combine the advantages in text and image modalities, and overcome the limitations in both modalities. It is important to find the relationships between the text and image modalities. Such relationships may be discovered via Latent Semantic Indexing, semantic networks, user's relevance feedback, and document clustering/classification.

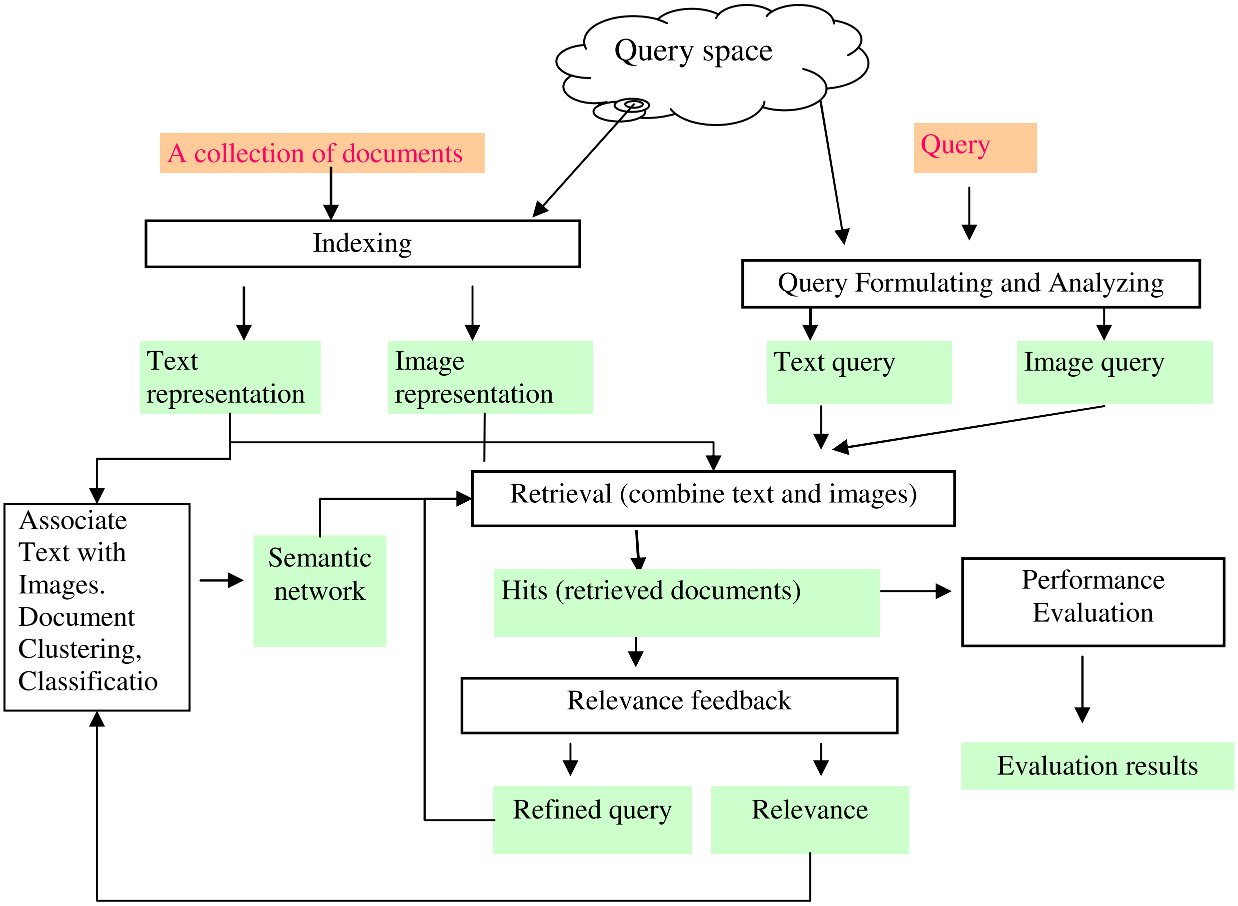


Figure 9. A diagram of a multimodal IR system.

Many research opportunities exist in multimodal IR. In the following, we discuss some open problems.

Dealing with imperfect data

When indexing multimodal documents, the data may be imperfect, such as collateral text in web pages, OCR results from document images, or noisy image segmentation results. When dealing with noisy data, indexing and retrieval techniques that can handle errors or competing interpretations are necessary.

Probabilistic integration models deserve further research since probability is a natural way of dealing with uncertainty. Integration of text and images using probabilistic retrieval models has been introduced in Section 4.3.3. These probabilistic models have a sound formalism. The relationships between text and images are automatically captured in the conditional probabilities of image given text or text given image. Variations of the probabilistic models have been proposed. The parameter estimations and the assumptions made are crucial for the effectiveness of the models. However, lack of standard data sets means it is difficult to demonstrate the power of these probabilistic models. Also the surveyed probabilistic models all use very simple multimodal documents, i.e. a document is an image with a set of keywords. There are challenges when generalizing the techniques based on simple documents to more complex multimodal documents. Also no surveyed probabilistic models incorporate user's feedback. How to integrate a user's feedback into the probabilistic model is an interesting issue.

Text and image combination techniques

Challenges exist in the combination of multimodal information. The data are heterogeneous and collateral text blocks have correlations with images. Issues that are worth further research include: how the text and image modalities can be optimally combined for a task without redundancy, how one modality can help the retrieval in the other modality, and how to intelligently combine the text and image queries to form hybrid queries.

The ideal combination is user specific. The users have diverse information needs and different users might use images differently. For example, Goodrum identifies two kinds of user's image use [Good00]: one is browsing and the other is search. Different modality might be used for different tasks. Browsing tasks may call for image attributes and visual examination of images of interest, while search tasks may require the specificity of text. Studies on user's information needs may be useful to identify at what point in their interaction with the retrieval systems users want or need to express a query using text, image or both. User centric retrieval relies on user's relevance feedback. User's relevance feedback techniques need further research.

Generality and Robustness

Techniques are needed to build a general and robust multimodal IR system. For a large quantity of multimodal documents, indexing all the features from multiple modalities is expensive. Most of the current systems use a static set of previously extracted features. It is a challenge to dynamically use only small parts of document data for indexing, to adapt to the changing needs of users and applications. For example, partial OCR results, partial layout analysis or partial figure analysis may be used for document image retrieval. Searching from a large set of heterogeneous multimodal documents and satisfying diverse user's queries are challenging.

A formal model for relationships between text and images

In the surveyed papers, ad-hoc experiments have shown that one modality can make another modality more informative and more precise, and the combination of two modalities usually improve performance. A formal model or formalism is needed to express the joint and disjoint information between images and text. That is, how much the inclusion of text can contribute to the improvement of image retrieval, how much the inclusion of images can contribute to the improvement of text retrieval, what can be achieved by the combination of text and images that is not possible with either alone, etc.

Using image for text tasks

The current research in image retrieval focuses on using collateral text to extract image semantics. The text retrieval research has not used images for text retrieval yet. For example, most of IR/IE tasks in biomedical literature use abstracts from MEDLINE [Hers04]. It is interesting to investigate how images in journal articles can be used to help text retrieval.

Performance evaluation

Multimodal document retrieval lacks standard benchmarks and datasets. When combining text and images, most of the systems report better performance. However, they use different data sets; it is hard to compare their methods. There is no standard comparison metrics. Currently, most researchers are using the standard evaluation metrics defined for text documents; these need to be extended or modified for multimodal documents. The Text Retrieval Conference (TREC) provides a common test platform for the researchers from text retrieval to evaluate their systems [VoHa98]. Experimentation is widely acknowledged as one of the driving forces behind the advancement of information retrieval. Systematic evaluation of the effectiveness of multimodal document retrieval is needed.

Borrow research achievements from video and other IR research areas

In our survey, we focus on the text and image modalities in multimodal IR. The techniques are mostly adapted from text retrieval or image retrieval. Multimodal indexing is also researched actively in video indexing, which have images, audio and text available [SnWo05]. The Informedia project has shown that automatic indexing of videos through the simultaneous analysis of both images and speech on the sound track can significantly improve indexing effectiveness [HaLK99]. Multimodal video indexing techniques may be borrowed to combine text and images. They are worth further research.

7. Acknowledgements

I gratefully acknowledge the advice from my Ph.D. supervisor, Dr. Dorothea Blostein, and my Ph.D. supervisor committee members, Dr. Hagit Shatkay and Dr. James Cordy. This research is supported by the Natural Sciences and Engineering Research Council of Canada.

Reference

Multimodal document retrieval combines techniques from different research areas. We use a short name in front of each reference to show which research area it belongs to. The meanings of the short names are as follows:
CL: Document classification/clustering; MM: Multimodal document retrieval; IM: image retrieval;
TR: text retrieval; DR: Document image retrieval; O: other.

IM [AnKJ02] S. Antani, R. Kasturi, R. Jain. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recognition, 35(2000), 945-965.

MM [BaFo01] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In ICCV, volume 2, pp. 408-415. IEEE Computer Society, 2001.

MM [BeCh02a] A.B. Benitez and S.F. Chang. Semantic knowledge constructions from annotated image collections. In Proceedings of the 2002 International Conference On Multimedia & Expo (ICME-02), Lausanne, Switzerland.

CL [BeCh02b] A.B. Benitez and S.F. Chang. Perceptual knowledge construction from annotated image collections. In Proceedings of the 2002 International Conference On Multimedia & Expo (ICME-02), Lausanne, Switzerland.

TR [BeLa99] A. Berger and J. Lafferty. Information retrieval as statistical translation. *SIGIR 22*, pp. 222–229. 1999.

MM [BeSC00] A.B. Benitez, J.R. Smith and S.F. Chang. MediaNet: a Multimedia Information Network for knowledge representation. IS&T/SPIE-2000, Vol. 4210, Boston, MA, Nov 6-8, 2000.

MM [BlJo03] D.M. Blei and M.I. Jordan. Modeling annotated data. SIGIR 2003.

O [Borg95] A. Borgida. Description logics in data management. IEEE Transactions on Data and Knowledge Engineering, 7:671-682,1995.

MM DR [BrCJ97] A. Bruce, V. Chalana, M.Y. Jaisimha, and T. Nguyen. The DocBrowse system for information retrieval from document image data. In Proc. Symposium on Document Image Understanding Technology, Annapolis, MD, 181-192.1997.

MM [BrZi04] D. Brahmi and D. Ziou. Improving CBIR systems by integrating semantic features. In Proc. RIAO, pages 291–305, Vaucluse, France, April 2004.

IM [CaTB99] C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, J. Malik. Blobworld: a system for region-based image indexing and retrieval. In Proc. of the Third International Conference on Visual Information and Information Systems (VISUAL'99), Appears in Lecture Notes in Computer Science, Vol. 1614, 1999, pp. 509–516.

DR [CeMS02] F. Cesarini, S. Marinai, and G. Soda. Retrieval by layout similarity of documents represented with MXY. Document Analysis Systems V (S. V.-L. 2423, ed.), pp. 353–364, 2002.

CL [ChBl04] N. Chen and D. Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. Accepted in December 2004, to be published by the International Journal of Document Analysis and Recognition.

CL [ChGN99] F. Chen, U. Gargi, L. Niles, and H. Schutz. Multi-modal browsing of images in web documents. In Proc. of SPIE Vol. 3651, p. 122-133. Document Recognition and Retrieval VI, Daniel P. Lopresti; Jiangying Zhou; Eds. 1999.

MM [ChLZ01] Z. Chen, W.Y. Liu, F. Zhang, M.J. Li and H.J. Zhang. Web mining for web image retrieval. Journal of the American Society for Information Science and Technology, 52(10), (2001), 831--839.

MM [ChSB97] S. Chang, J. Smith, M. Beigi & A. Benitez. Visual information retrieval from large distributed online repositories. Communications of ACM 40, 63-71.

IM [ChSY87] S. K. Chang, Q. Y. Shi, and C. Y. Yan. Iconic indexing by 2-D strings. IEEE Trans. on Pattern Anal. Machine Intell., Vol.9, No.3, pp. 413-428, May 1987.

TR [CrLR98] F. Crestani, M. Lalmas, C. J. van Rijsbergen, and I. Campbell. Is this document relevant? ... Probably: a survey of probabilistic models in information retrieval. ACM Computing Surveys 30(4): 528–552. 1998.

CL [CuKP92] D.R. Cutting, D.R. Karger, J.O. Pedersen and J.W. Tukey. Scatter/Gather: a cluster-based approach to browsing large document collections. In Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), 318-329, 1992.

TR [DeDF90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, Volume 41, Issue 6, 1990.

O [DeLR77] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, series B, 39(1):1–38, 1977.

DR [Doer98] D. Doermann. The indexing and retrieval of document images: a survey. Computer vision and image understanding, 70(3): 287-298. 1998.

CL [DuBF02] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Proc. of the Seventh European Conference on Computer Vision, pages 97–112. 2002.

MM [Duff98] G. Duffing. Text-image interaction for image retrieval and semi-automatic indexing. In Proc. of the 20th Annual BCS-IRSG Colloquium on IR, 1998.

IM [Flic95] M. Flickner et al. Query by image and video content: the QBIC system. IEEE Computer, pp. 23–30, Sep. 1995.

CL [GeAl03] T. Gevers and F. Aldershoff. Classifying multimedia documents by merging textual and pictorial information. ICIP (3) 2003: 13-16.

MM [GoCL04] K. Goh, E.Y. Chang and W.C. Lai. Multimodal concept dependent active learning for image retrieval. MM'04, October 10-16, 2004, New York, New York, USA.

IM [Good00] G. Goodrum. Image information retrieval: an overview of current research. Information Science, Vol. 3, No.2 , 2000.

IM [GuRa95] V. N. Gudivada and V. Raghavan. Design and evaluation of algorithms for image retrieval by spatial similarity. ACM Transactions on Information Systems, 1995, April, Vol.13, No. 2, pp.115-144.

MM [HaLK99] A.G. Hauptmann, D. Lee, and P.E.Kennedy. Topic labeling of multilingual broadcast news in the informedia digital video library. In Proc. of ACM DL/SIGIR MIDAS Workshop, Berkely, USA, 1999.

IM [HaSD73] R. M. Haralick, K. Shanmugam, and I. Dinstein. Texture features for image classification. IEEE Trans. On Sys. Man. and Cyb. SMC-3(6), 1973.

IR [Hers04] W. R. Hersh, et al. TREC 2004 genomics track overview. The Thirteenth Text Retrieval Conference: TREC 2004. 2004. Gaithersburg, MD: National Institute of Standards and Technology.

TR [Hiem98] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. ECDL 2, pp. 569–584. 1998.

CL [Hofm98] T. Hofmann. Learning and representing topic, a hierarchical mixture model for word occurrence in document databases. Workshop on learning from text and the web. 1998.

IM [HuKM97] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu and R. Zabih. Image indexing using color correlograms. In Proc. IEEE Conference on CVPR, (1997) 762—768.

IM [HuMR96] T. S. Huang, S. Mehrotra, and K. Ramachandran. Multimedia analysis and retrieval system (MARS) project. In Proc. of 33rd Annual Clinic on Library Application of Data Processing-Digital Image Access and Retrieval,1996.

MM [Ingw96] P. Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. Journal of Documentation 52(1), 3-50. 1996.

CL [JaMF99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: a review. ACM Computing Surveys, 1999.

IM [JaVa98] A.K. Jain and A. Vailaya. Shape-based retrieval: a case study with trademark image databases. Pattern Recognition 31 (9) (1998) 1369–1390.

MM [JeLM03] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 119–126, July 28-August 1, 2003.

O [KaTC02] J. Kandola, J. Shawe-Taylor, N. Cristianini. Learning semantic similarity. NIPS (2002).

TR [LaCC02] V. Lavrenko, M. Choquette, and W. Croft. Cross-lingual relevance models. In Proceedings of the 25th annual international ACM SIGIR conference, pp.175–182, 2002.

TR [LaZh01] J. Laferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual international ACM SIGIR Conference, pp.111-119, 2001.

MM
DR [LoMo95] O. Lorenz and G. Monagan. A retrieval system for graphical documents. In Symposium on Document Analysis and Information Retrieval, pp. 291-300, 1995.

DR [Lopr96] D. P. Lopresti. Robust retrieval of noisy text. In Proc. of ADL'96, pp. 76–85, 1996.

CL [LuDr01] C. Lu, M. Drew. Construction of a hierarchical classifier schema using a combination of text-based and image-based approaches. In Proc. of SIGIR'01, September 9-12, Louisianna, USA.

MM [LuZH03] Y. Lu, H. Zhang, W.Y. Liu and C. Hu. Joint semantics and feature based image retrieval using relevance feedback. IEEE Transactions on Multimedia, 5(3):339-347, September 2003.

DR [LuZT04] Y. Lu, L. Zhang, C.L.Tan. Retrieving imaged documents in digital libraries based on word image coding. In Proc. of the first international workshop on Document Image Analysis for Libraries (DIAL'04).

IM [Mall89] S. G. Mallat. A theory for multi resolution signal decomposition: the wavelet Representation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 11, pp. 674-693, July 1989.

IM [MeKL97] B. M. Mehtre, M. Kankanhalli, and W. F. Lee. Shape measures for content based image retrieval: A comparison. Information Processing & Management 33(3), 1997.

MM [MeSS01] C. Meghini, F. Sebastiani, U. Straccia. A model of multimedia information retrieval. Journal of ACM, 48(5):909-970, 2001.

DR [MiCh00] M. Mitra and B.B. Chaudhuri. Information retrieval from documents: a survey. Information retrieval, vol. 2, nos. 2/3, pp.141-163, 2000.

TR [Mill95] G.A. Miller. WordNet: a lexical database for English. Comm. of the ACM, Vol. 38, No. 11, pp. 39-41, Nov.1995.

TR [MiLS99] D.R.H. Miller, T. Leek, and R.M. Schwartz. A Hidden Markov Model information retrieval system. SIGIR 22, pp. 214–221. 1999.

CL [MiPi95] R.W. Picard and T. P. Minka. Vision texture for annotation. Multimedia Systems, 3(1):3–14, 1995.

MM [MuCh99] S. Mukherjea and J. Chost. Automatically determining semantics for World Wide Web multimedia information retrieval. Journal of Visual Languages and Computing (1999) 10, 585-606.

MM [NaKT03] A. Nakagawa, A. Kutics, K. Tanaka, and M. Nakajima. Combining words and object-based visual features in image retrieval. In Proc. of the 12th International Conference on Image Analysis and Processing (ICIAP'03), pp. 354-359, September 2003.

DR [NaSe84] Nagy and S. Seth. Hierarchical representation of optically scanned documents. In Proceedings of the 7th international conference on pattern recognition, Los Alamitos, California, USA, 1984, pp 347-349.

MM [OgSt95] V. E. Ogle and M. Stonebraker. Chabot: retrieval from a relational database of images. IEEE Computer, Vol. 28, No. 9 (Sept. 1, 1995): 40-48.

MM [OrPM99] M. Ortega, K. Porkaew and S. Mehrotra. Information retrieval over multimedia documents. In Proc. of 1999 SIGIR post-conference workshop on Multimedia Indexing and Retrieval, Berkeley, CA, August,1999.

CL [PaSH99] S. Paek, C. Sable, V. Hatzivassiloglou, A. Jaimes, B. Schiman, S. F. Chang, and K. McKeown. Integration of visual and text-based approaches for the content labeling and classification of photographs. In Proceedings of the ACM SIGIR Workshop on Multimedia Indexing and Retrieval (SIGIR-99).

MM [PaSm98] S. Paek and J.R. Smith. Detecting image purpose in World Wide Web documents. Trans. Patt. Analys. Mach. Intell. 22, 12, 1349–1380. 1998.

IM [PeFu77] E. Persoon and K. S. Fu. Shape discrimination using Fourier descriptors. IEEE Trans. Sys. Man. Cyb. 1977.

TR [PoCr98] J.M. Ponte and W.B. Croft. A language modeling approach to information retrieval. In Proc. of SIGIR 21. 1998.

TR [Rijs79] C. J. van Rijsbergen. Information Retrieval, on-line book. 1979.

TR [Rocc71] J. J. Rocchio. Relevance feedback in information retrieval. In The SMART Retrieval System -- Experiments in Automatic Document Processing, pp. 313-323, Englewood Cliffs, NJ, 1971. Prentice Hall, Inc.

TR [RoSp76] S. Robertson and K. Sparck Jones. Relevance weighting of search terms. Journal of the American Society for Information Science, 27:129–146. 1976.

MM [Rowe 95] N.C. Rowe. Retrieving captioned pictures using statistical correlations and a theory of caption-picture co-reference. In Symposium on Document Analysis and Information Retrieval, pages 525-534, 1995.

MM [Rowe 99] N.C. Rowe. Precise and efficient retrieval of captioned images: the MARIE project. Library Trends, Fall 1999.

IM [RuHC97] Y. Rui, T. S. Huang, and S. Chang. Image retrieval: past, present and future. Int. Symposium on Multimedia Information Processing, Dec 11-13, 1997, Taipei, Taiwan.

IM [RuHu99] Y. Rui and T. S. Huang. Image retrieval: current techniques, promising directions, and open issues. Journal of Visual Communication and Image Representation 10,39-62(1999).

TR [SaBu90] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4):288-297, 1990.

TR [Salt89] G. Salton. Automatic text processing. Addison-Wesley, 1989.

IM [Same84] H. Samet. The Quadtree and related hierarchical data structures. ACM Computing Surveys, Vol.16, No.2, pp.187-260, 1984.

MM [Sant02] S. Santini. Multimodal search in collections of images and text. Journal of Electronic Imaging, Volume 11, Issue 4, pp. 455-468, 2002.

TR [SaWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620.1975.

MM [ScCS99] S. Sclaroff, M. La Cascia, and S. Sethi. Unifying textual and visual cues for content-based image retrieval on the world wide web. Comp. Vis. IU, 75:86-98, 1999.

IM [ScTC97] S. Sclaroff, L. Taycher, and M. La Cascia. ImageRover: A content-based image browser for the world wide web. In Proc. of IEEE Int. Workshop on Content-Based Access of Image and Video Libraries, 1997.

CL [Seba02] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys 34(1): 1-47. 2002.

CL [ShDR01] C. Shin, D. Doermann, A. Rosenfeld. Classification of document pages using structure-based features. International Journal on Document Analysis and Recognition 3(4): 232-247

DR [SmSp97] A. F. Smeaton and A. L. Spitz. Using character shape coding for information retrieval. In Proc. of 4th ICDAR, pp. 974–978, 1997.

MM [SnWo05] C. Snoek and M. Worring. Multimodal video indexing: a review of the state-of-the-art. Multimedia Tools and Applications, 25, 5–35, 2005.

IM [SmWS00] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

IM [SqMP00] D. M. Squire, W. MTuller, H. MTuller, T. Pun. Content-based query of image databases: inspirations from text retrieval. Pattern Recognition Letters (Selected Papers from The 11th-Scandinavian Conference on Image Analysis SCIA '99) 21 (13-14) (2000) 1193-1198.

MM [SrCB94] R. K. Srihari, R. Chopra, D. Burhans, M. Venkataraman, and V. Govindaraju. Use of collateral text in image interpretation. In Proc. of ARPA Image Understanding Workshop 1994.

MM [Srih00] R.K. Srihari. Intelligent indexing and semantic retrieval of multimodal documents. Information Retrieval, 2, 245-275, 2000.

IM [StOr95] M. Stricker and M. Orengo. Similarity of color images. In Proc. of SPIE Storage and Retrieval for Image and Video Databases, 1995.

MM [SwFA97] M. Swain, C. Frankel, and V. Athitsos. Webseer: an image search engine for the world wide web. CVPR, 1997.

DR [TaBC94] K. Taghva, J. Borsack, A. Condit, and S. Erva. The effects of noisy data on text retrieval. Journal of the American Society for Information Science, 45:50-58, 1994.

IM [TaMY78] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. IEEE Trans. On Sys., Man. and Cyb. SMC-8(6), 1978.

IM [VaLi00] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2000), IEEE Computer Society, Hilton Head Island, South Carolina, USA, 2000, pp. 216-221.

IM [Vasc00] N. Vasconcelos. Bayesian models for visual information retrieval. PhD thesis, Massachusetts Institute of Technology, 2000.

IM [VeHa01] R. C. Veltkamp and M. Hagedoorn. State-of-the-art in shape matching. In Michael Lew, editor, Principles of Visual Information Retrieval, pages 87–119. Springer, 2001. ISBN 1-85233-381-2.

IM [VeTa02] R. C. Veltkamp and M. Tanase. Content-based image retrieval systems: a survey. Revised and extended version of Technical Report UU-CS-2000-34, October 2002.

TR [VoHa98] E.M. Voorhees and D. Harman. Overview of the sixth Text REtrieval Conference (TREC-6). In Proceedings of the Sixth Text REtrieval Conference(TREC-6), August 1998. pp.1-24.

TR [Voor94] E.M. Voorhees. Query expansion using lexical-semantic relations. In Proc. of the 17th International Conference on Research and Development in Information Retrieval SIGIR'94 Dublin, Ireland. 1994.

CL [WaLW01] J. Z. Wang, J. Li, G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. IEEE Transactions on pattern Analysis and Machine Intelligence 23 No 9 (2001) 1-17.

MM [WaMX04] X.J. Wang, W.Y.Ma, G.R.Xue, X.Li. Multi-modal similarity propagation and its application for web image retrieval. MM'04, Oct. 10-16, NewYork, NY, USA.

MM [West00] T. Westerveld. Image retrieval: Content versus context. In Proc. of Computer-Assisted Information Retrieval, Vol. 1, Paris, France, 2000, pp. 276-284.

CL [Will88] P. Willet. Recent trends in hierarchical document clustering: A critical review. Information Processing and Management, 24:577-597, 1988.

TR [Witt98] I. H. Witten, et al. Managing Gigabytes, compressing and indexing documents and images (2nd ed.), Morgan Kaufmann, San Diego, CA. 1999.

IM [WoSS00] M. Worring, A. W. M. Smeulders, S. Santini. Interaction in content-based image retrieval: An evaluation of the state of the art. In: R. Laurini Ed.), Fourth International Conference On Visual Information Systems (VISUAL'2000), no. 1929 in Lecture Notes in Computer Science, Springer-Verlag, Lyon, France, 2000, pp. 26-36.

TR [YaNe99] R. B.Yates and B. R. Neto. Modern information retrieval. Addison-Wesley/ACM Press, 1999.

IM [YoIc99] A. Yoshitaka and T. Ichikawa. A survey on content-based retrieval for multimedia databases. IEEE Trans. Knowledge and Data Eng., vol. 11, no. 1, Jan./Feb. 1999, pp.81–93.

MM [ZhGr02] R. Zhao and W. I. Grosky. Narrowing the semantic gap-improved text-based web document retrieval using visual features. IEEE Transactions on Multimedia, 4(3):189-200, June 2002.

MM [ZhHu02] X. S. Zhou and T. S. Huang. Unifying keywords and visual contents in image retrieval. IEEE Multimedia, 4(2):23-33, June 2002.

IM [ZhHu03] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 8: 536–544 (2003).

MM [ZhSu03] H.Z. Zhang and Z. Su. Relevance feedback and learning in content-based image search. WWW: Internet and Web Information Systems, 6(2):131–155, 2003.

MM [ZhZO04] Z. Zhang, R. Zhang, J. Ohya. Exploiting the cognitive synergy between different modalities in multimodal information retrieval. 2004 IEEE International Conference on Multimedia and Expo (ICME),2227-2230.