Extracting Bimodal Representations for Language-Based Image Retrieval[1]

Thijs Westerveld, Djoerd Hiemstra, Franciska de Jong

University of Twente, Centre for Telematics and Information Technology,
PO Box 217, 7500 AE Enschede, The Netherlands
{westerve, hiemstra, fdejong}@cs.utwente.nl

Abstract This paper explores two approaches to multimedia indexing that might contribute to the advancement of text-based conceptual search for pictorial information. Insights from relatively mature retrieval areas (spoken document retrieval and cross-language retrieval) are taken as a starting point for an investigation of the usefulness of the concept of bimodal dictionaries and of clustering features from multi-modal documents into one semantic space. One of the advantages of the presented techniques is that they are domain independent.

1 Introduction

Among the various types of objects that one could want to search for in the multimedia domain, image content seems to be one of the more challenging types. Speedy and easy access to image content in the general domain is not supported by today’s search tools and technology, and in spite of progress in content based image retrieval or advances in the area of video logging (a technique which reuses subtitles or speech transcripts for the automatic indexing of video fragments [13]), it is unlikely that on short notice general purpose technology will become available for indexing images on a conceptual level.

In this paper we will investigate parallels for image content retrieval with retrieval areas that have already reached a certain level of maturation (text retrieval), or where a breakthrough is imminent (spoken document retrieval and cross-language retrieval) and where the incorporation of the available technology in digital workflow processes can already be shown at several places.

In order to address the parallelism at a proper level of abstraction, section 2 will first discuss the retrieval problem in a very general way, based on the classical paradigms for information retrieval. Then we will analyse two more complex retrieval tasks. In both cases the complexity is not matching the complexity of the image case, but still it is complex enough to be able to generalise the findings in such a way that it offers a useful baseline for a conceptual exploration of the problem of developing automatic indexing tools for the image domain. In section 3 the notion of deriving

[1] Part of the research for this paper has been funded by the Dutch organisation for Scientific Research NWO.

bimodal dictionaries from parallel corpora will be shown to be very useful in modelling an approach for image retrieval which will eventually support text based querying of images. In section 4 the construction of one semantic space for both text and image features via *Latent Semantic Indexing* will be shown to offer a useful basis for concept-based retrieval.

Research in computer vision can be argued to advance the search for pictures in comparison to text-based search using manual annotations to pictures [15]. The research described in this paper explores the possibility to reuse *all* textual material that accompanies a picture for the purpose of automatic indexing. The concluding section 5 will argue that because of the parallelism with more matured retrieval technology and because of the fact that textual queries can be linked to the conceptual level there is room for confidence that the approaches described may facilitate a next advancement in cross-modal retrieval.

2 Learning from other IR areas

The goal of an information retrieval system is to present the user a set of documents that satisfy the user's information need which is expressed by a query. Every approach to information retrieval has three basic components: (1) a method for representing documents, (2) a method for representing queries and (3) a method for comparing the two representations [14]. Figure 1 graphically displays the basic components of an IR system.

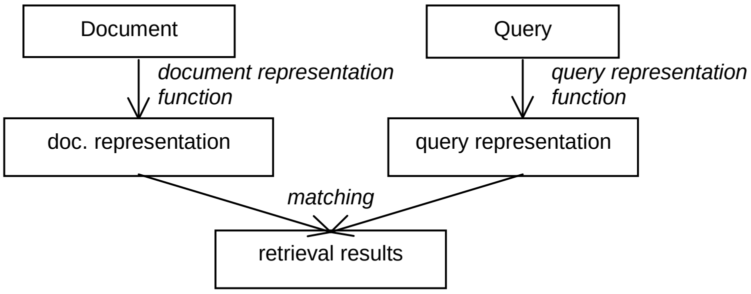


Fig. 1. The information retrieval paradigm

In the case of textual documents in a specific language and textual queries in the same language, both the query and document representation functions can be kept relatively simple. Still, an important problem in text retrieval is the fact that the same concept can be formulated in many different ways. Because of that, it is very hard to predict which query will yield relevant documents on a given concept. People search for documents discussing certain issues and are less interested in the specific words used in the documents to convey this information. This problem is known as the paraphrase problem [14]. Cross-modal IR, cross-language IR and spoken document retrieval can be viewed as special instances of the paraphrase problem. Queries and documents necessarily use different vocabularies because of differences in language or modality.

The following sections will go deeper into approaches and resources for cross-language information retrieval (CLIR) and spoken document retrieval (SDR).

2.1 Approaches to CLIR and SDR

A CLIR system supports the use of textual queries to search documents in other languages. A SDR system supports the use of textual queries to search audio documents containing speech. In systems that support CLIR or SDR, either the document representation function or the query representation function (or both) need to be much more complicated than in the classical text retrieval case.

In the case of CLIR the system can either translate all of the documents to the language of the query (off-line document translation) [10], or translate the query to the language of the documents (query translation) [8], or translate both to some intermediate language. Document translation has the advantage that it can be done off-line during indexing. Furthermore, the quality of document translation will in principle be better because the full document context is available for disambiguation. In the case of query translation there is often very little context. Latent semantic indexing (LSI) is an approach to CLIR in which both documents and queries are represented in a format that is in some sense intermediate between the two languages [4,16].

In the case of SDR, one option is to apply large-vocabulary speech recognition on the spoken documents to generate a textual representation [13]. Another option is phone-based retrieval. Just like in speech recognition, the system extracts discrete features from the continuous speech signal: so-called phones, but the final recognition step in omitted. During searching the query is also converted into a phone representation [11]. Phone-based retrieval is much faster during indexing and is less sensitive to out of vocabulary words.

2.2 Resources for CLIR and SDR

Both CLIR and SDR may rely on research areas that have proven to be commercially successful: machine translation and speech recognition. In addition, CLIR systems might also be built using machine-readable human dictionaries [8], parallel corpora (two texts that are translations of each other) [16] or comparable corpora (texts in two languages on the same subject which are not direct translations) [1]. SDR systems typically use speech recognition software [13], but might also use pronunciation dictionaries and transcribed speech, which are used to train the speech recognition software, directly [11].

2.3 Language-based image retrieval

In the sections above a brief overview was given of approaches and resources for CLIR and SDR. Drawing the parallel with language-based image retrieval, three basic approaches can be identified: (1) off-line image recognition, (2) on-line ’query visualisation’ and (3) the use of an intermediate representation. If the available resources are taken into account it becomes apparent that the first option will not be realised in the near future. Unlike machine translation systems for CLIR and speech recognition systems for SDR, the state of the art in computer vision does not permit the automatic full interpretation of an arbitrary scene [15].

However, there are annotated image collections. In a sense these resources are resembling manually produced dictionaries for CLIR and pronunciation dictionaries for SDR. Such collections can be searched using textual queries, but new pictures should always be annotated by hand.

Systems that support CLIR or SDR can be trained on parallel corpora and transcribed speech respectively to interpret new or unseen data. We argue that the same applies to language-based image retrieval if large image/text corpora are available. Such corpora might be acquired from the Internet by using complete web pages plus images or e.g. from newspapers by using complete articles plus corresponding photos. This paper suggests two approaches to language-based image retrieval using large mixed image/text corpora. The first approach is the automatic extraction of a bimodal dictionary from the corpus. The second approach is based on clustering of both text and image features into one conceptual space using LSI.

3 Extracting a Bimodal Dictionary from a Parallel Corpus

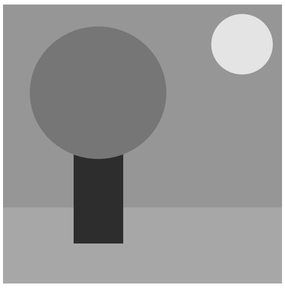
If we have a parallel corpus of documents in the modalities text and images (i.e. a set of pictures with reasonably complete descriptions), we could extract a bimodal 'translation' lexicon from this corpus using techniques developed for cross language information retrieval [9]. Such a translation lexicon can be extracted by looking at the number of times language elements and image elements co-occur. If this number is relatively high, the elements are probably related. For example, if image descriptions contain the word *sun* and the corresponding images often contain a yellow circle, we could conclude that the two elements (the word *sun* and a yellow circle) are related in a way that is similar to the relation between *sun* and *soleil.*

3.1 Pre-processing

In order to be able to extract such relations from a parallel corpus of images and text, both the images and the descriptions need to be pre-processed.

The pre-processing of images consists of the determination of the objects or features present. In practice, low-level features like e.g. colour-histograms, textures and hue and saturation values are used to describe images. Low-level image features proofed to be useful for image retrieval systems that use query-by-example [5,7,12]. For the sake of our example, we assume that some basic shapes and colours like e.g. 'yellow circle' are identified.

For the pre-processing of textual objects we assume the application of a statistical part-of-speech tagger for lemmatising the words in the text, thereby following the



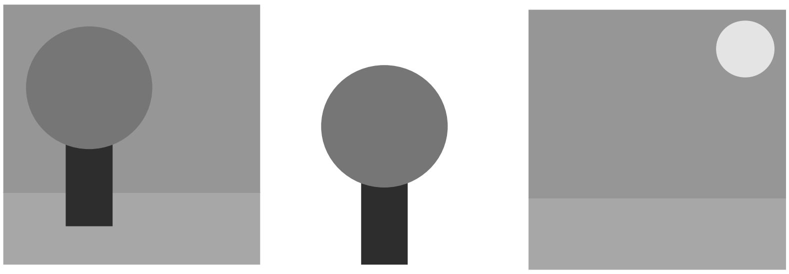
A tree in the park on a sunny day

Fig. 2. Example image and description

approach taken in NLP-enhanced retrieval systems such as Twenty-One [10]. Suppose we have a picture of a tree in a park with the accompanying description “A tree in the park on a sunny day” (**Fig. 2**). The features detected for this image might be: a green circle, a yellow circle, a brown rectangle, a green rectangle and a blue rectangle. The information elements from the text might be *tree*, *park*, *sun* and *day*.

3.2 Counting Co-occurrences

Once the images and descriptions have been pre-processed, we can start counting the number of co-occurrences for the various text and image elements. Suppose we have a small example corpus, consisting of three pictures on the one hand, and corresponding descriptions on the other hand. (**Fig. 3**).



A **tree** in the **park**

A **tree**

A **sun**ny **day** in the **park**

Fig. 3. Example corpus[2]

To compute co-occurrence frequencies from the corpus, a table will be constructed in which the columns represent the words in the captions, whilst the rows represent the features in the pictures. Since the mapping between words and image features will not always be one-to-one, the number of words in a description might differ from the number of features in the corresponding image. When this happens, null-terms will be introduced, to balance the two.

Initially all cells in the table will be set to zero. Each time a word and a feature co-occur, the corresponding cell in the table will be increased. The counts for the co-occurrences are distributed equally over all possibilities. For example, in the last image, *yellow circle* might be a translation of either, *park*, *sun* or *day*, so the three cells corresponding to these translations are all increased by 0.33.

When all three pictures from our example corpus have been processed, the table looks like this:

	tree	park	sun	day	(null)
green circle	0.75	0.25	-	-	1.0
yellow circle	-	0.33	0.33	0.33	-
brown rectangle	0.75	0.25	-	-	1.0
blue rectangle	0.25	0.58	0.33	0.33	0.5
green rectangle	0.25	0.58	0.33	0.33	0.5

[2] The information elements in the captions are in bold face

The following bimodal dictionary can be inferred automatically by applying the EM-algorithm as described in [9]. The algorithm eliminates certain word image-feature combinations and infers for example that trees basically consist of green circles and brown rectangles.

	tree	park	sun	day	(null)
green circle	1.0	-	-	-	1.0
yellow circle	-	-	0.5	0.5	-
brown rectangle	1.0	-	-	-	1.0
blue rectangle	-	1.0	0.25	0.25	0.5
Green rectangle	-	1.0	0.25	0.25	0.5

The algorithm that is used to infer the bimodal dictionary can easily be generalised for non-discrete lower level features. Future experiments have to point out if in fact the resulting bimodal dictionaries are effective in cross-modal search situations.

4 Using LSI for Cross-Modal Information Retrieval

Another way to find relations (or translations) between words and image features is building a semantic space with LSI. The idea of using LSI for multi-modal indexing is not completely new. In another study [2], images from the world wide web were indexed, analysing both the image features and the surrounding text (alt-tags, title, near text) of an image. In this study, words and image features are analysed separately building two separate spaces that are merged afterwards. LSI is used to analyse the textual part. A more integrated approach, in which the two modalities are combined from the start will yield a semantic space in which text and images are firmly interwoven.

As stated above, LSI can be used for Cross-Language Information Retrieval. After a short introduction to LSI this section will explain how the same techniques can be used for Cross-Modal Retrieval. For a more detailed description of LSI, cf. [3].

4.1 Latent Semantic Indexing

LSI uses a matrix in which each term is represented as a vector of document occurrences. The elements in the vector can be binary, natural or real-valued, indicating either *if* a term occurs, *how often* a term occurs or what the *weight* of the occurrence is (e.g. tf/idf).

From this matrix, the most meaningful linear combinations of terms and documents are computed using Singular Value Decomposition (SVD) [6], a form of two-mode factor-analysis that is similar to eigen-value/eigen-vector decomposition. The original term-document-matrix A is decomposed into three matrices containing linearly independent components that together represent the original term document matrix:

$$A = T\ S\ D^t \qquad (1)$$

T and D are matrices with left and right-singular vectors, representing respectively terms and documents. S is a diagonal matrix with singular values that are ordered in decreasing magnitude.

When the smaller singular values in S are ignored (i.e. set to zero), the product of the three matrices is an optimal approximation of the original matrix in a lower dimension. In this lower dimensional space, documents that contain different, but similar[3] terms are mapped onto the same vector. This way, the underlying, latent semantics of the documents are uncovered. The similarity between two documents is computed using the cosine measure of the transformed documents. The documents are transformed using the singular value matrices:

$$\mathrm{sim}(d_1,d_2) = \cos(T^t d1, T^t d_2) \qquad (\mathbf{2})$$

where T is the matrix with the singular vectors representing the terms.

4.2 Building a Cross-modal Semantic Space

To use LSI for cross-modal indexing and retrieval, a term-document matrix has to be constructed. Analogous to cross-language retrieval with LSI, documents from the two modalities are treated as the same document. So along the rows of the matrix we will have both text and image features.

Suppose we use the example corpus from **Fig. 3** to build the matrix, then the matrix will look like this:

		doc 1	doc 2	doc 3
Terms from the text:	tree	1	1	0
	park	1	0	1
	sun	0	0	1
	day	0	0	1
Terms from the images:	green circle	1	1	0
	brown rectangle	1	1	0
	blue rectangle	1	0	1
	green rectangle	1	0	1
	yellow circle	0	0	1

Singular Value Decomposition is used to decompose this matrix into three matrices of a lower dimensionality. The semantic space that is constructed from this small example corpus via reduction to two dimensions is shown in **Fig. 4.**

In a way, the semantic space can be regarded as a bimodal dictionary. In this space, related terms from the two modalities (translations) are located close to each other.

[3] similar terms is, in this case, terms that often occur together in documents.

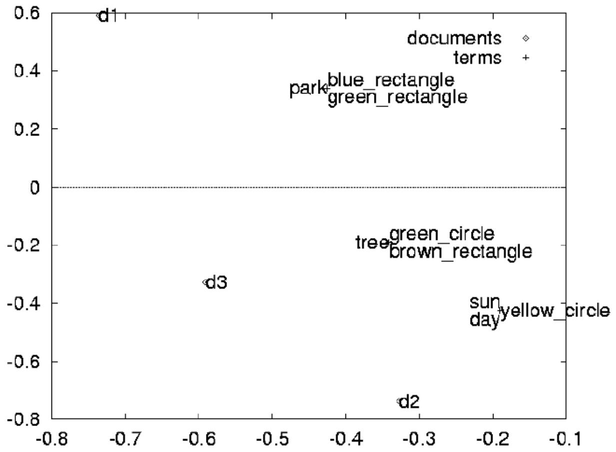


Fig. 4. 2-dimensional semantic space of example corpus from **Fig. 3**

4.3 Using a mixed corpus

Application of the described approach is hindered by the fact that parallel corpora of text and images are not widely available. However, it is also possible to build a similar semantic space using a mixed corpus. A mixed corpus is a corpus consisting of documents that contain text or images or both (**Fig. 5**). In a mixed corpus, at least a number of documents need to contain both text and images, but the text and the images don't need to be translations of one another, they only need to be *about* the same subject. Mixed corpora are much easier to obtain. We could take for example documents from the World Wide Web.

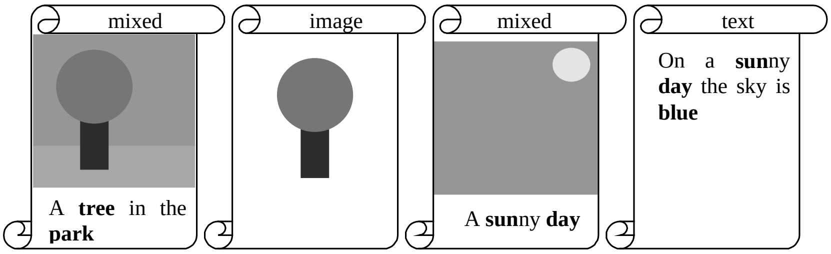

mixed
A **tree** in the **park**
image
mixed
A **sun**ny **day**
text
On a **sun**ny **day** the sky is **blue**

Fig. 5. Mixed corpus

Once we have a corpus, we can apply proven techniques developed for conceptual monolingual information retrieval to build multi-modal and even cross-modal indices.

In concept-based retrieval, the assumption is that documents can communicate the same information in various wordings. The same assumption can be made on multi-modal documents. These documents can communicate the same information using different words, different images or even words instead of images and vice versa.

So, from a mixed corpus, again, a term-document matrix can be constructed in which *terms* from both text and images are represented. This matrix can then be decomposed using SVD and in the resulting space related terms will be close to each other. Both textual and visual queries, as well as multi-modal ones, can be mapped onto this space to find documents about a specific concept in any modality.

5 Concluding Remarks

In this paper, we have addressed the problem of cross-modal and multi-modal indexing. We have shown that multi-modal indexing can be regarded as just another case of multi-lingual indexing. Just like in cross language information retrieval and in spoken document retrieval, we are dealing with a set of documents and queries that cannot be compared in a trivial way because they differ in language or modality.

In order to be able to retrieve documents in one modality using a query in another modality, we need a suitable mechanism to map both query and documents to a common representation language. We have presented two methods for mapping documents in different modalities to the same common language:

1. deriving a bimodal dictionary
2. multi-modal indexing using LSI

Both methods use the fact that the co-occurrence of two terms (visual or textual) in a text is not purely coincidental. Both methods automatically derive the relationships between terms from a parallel (1) or mixed (2) corpus. Therefore, both methods heavily depend on the availability of such a corpus.

Since, corpus based methods have proven to be successful in both CLIR and SDR, and indexing images based on low-level image features works fairly well, we are confident that extracting bimodal representations from multi-media corpora may be useful for language based image indexing and retrieval.

References

1. Brachsler, M. and Schäuble, P. *Multilingual Information Retrieval Based on Document Alignment Techniques,* Proceedings of the second European Digital Libraries Conference, 1998.
2. Cascia, M. La, Sethi, S., Sclaroff, S., *Combining textual and visual cues for content-based image retrieval on the world wide web,* In: IEEE Workshop on content-based access of image and video libraries, 1998
3. Deerwester, S., Dumais, S.T., Harshman, R., *Indexing by Latent Semantic Analysis,* In: Journal of the American Society for Information Science, 41 (6), pp391-407, 1990.
4. Dumais, S.T., Landauer, T.K., Littman, M.L., Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing, In: Proceedings SIGIR96, Workshop On Cross-Linguistic Information Retrieval, 1996.

5. Flickner, M., Sawhney, H., Niblack, W. Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D. Steele, D. and Yanker, P. *Query by image and video content: the QBIC system*, In: Maybury, M.T. (ed.) Intelligent multimedia information retrieval, pages 7-22, 1997.
6. Forsythe, G.E., Malcolm, M.A. and Moler, C.B., *Least squares and the singular value decomposition* In: Computer Methods for Mathematical Computations, (Chapter 9), Englewood Cliffs, NJ: Prentice Hall, 1977.
7. Gevers, T. and Smeulders, A.W.M., *PicToSeek: A content-based image search engine for the WWW*, In: Proceedings of VISUAL'97, 1997
8. Hiemstra, D. and W. Kraaij, *Twenty-One at TREC-7: Ad-hoc and Cross-language track*¸ In: Proceedings of the seventh Text Retrieval Conference TREC-7, Nist Special Publications, 1999.
9. Hiemstra, D., *Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus*, In: Peter-Arno Coppen, Hans van Halteren and Lisanne Teunissen (eds.), Proceedings of the eighth CLIN meeting, pages 41-58, 1998.
10. Jong, F. de, *Twenty-One: a baseline for multimedia retrieval*, In: Proceedings of the 14th Twente Workshop on Language Technology TWLT-14, pp 189-195,1998.
11. Kraaij, W., J. van Gent, R. Ekkelenkamp, and D. van Leeuwen, *Phoneme-based Spoken Document Retrieval*, In: Proceedings of the 14th Twente Workshop on Language Technology TWLT-14, pp 141-153, 1998.
12. Marsicoi, M., Clinque, L. and Levialdi, S., *Indexing pictorial documents by their content: a survey of current techniques*, In: Image and vision computing 15, pages 119-141, 1997
13. Netter, K. and F. de Jong, *Olive: speech based video retrieval*, In: Proceedings of the 14th Twente Workshop on Language Technology TWLT-14, pp 187-189, 1998.
14. Oard, D.W. and Dorr, B.J., *A Survey of Multilingual Text Retrieval*, Technical report TR-96-19, University of Maryland, http://www.ee.umd.edu/medlab/mlir/mlir.html
15. Smeulders, A.W.M., T. Gevers and M.L. Kersten, *Computer vision and image search engines.* In: Proceedings of the 14th Twente Workshop on Language Technology TWLT-14, pp 107-116, 1998.
16. Yang, Y. Carbonell, J.G., Brown, R.D., Frederking, R.E., *Translingual Information Retrieval: Learning from Bilingual Corpora,* Artificial Intelligence 103 (1-2), pp 323-345, 1998.