Multimodal Information Spaces for Content-based Image Retrieval

Juan C. Caicedo
National University of Colombia
<http://www.informed.unal.edu.co>
jccaicedoru@unal.edu.co

Abstract

Currently, image retrieval by content is a research problem of great interest in academia and the industry, due to the large collections of images available in different contexts. One of the main challenges to develop effective image retrieval systems is the automatic identification of semantic image contents. This research proposal aims to design a model for image retrieval able to take advantage of different data sources, i.e. using multimodal information, to improve the response of an image retrieval system. In particular two data modalities associated to contents and context of images are considered in this proposal: visual features and unstructured text annotations. The proposed framework is based on kernel methods that provide two main important advantages over the traditional multimodal approaches: first, the structure of each modality is preserved in a high dimensional feature space, and second, they provide natural ways to fuse feature spaces in a unique information space. This document presents the research agenda to build a Multimodal Information Space for searching images by content.

1. INTRODUCTION

Content-Based Image Retrieval (CBIR) is an active research discipline focused on computational strategies to search for relevant images based on visual content analysis. In this proposal, multimodal analysis is considered to develop CBIR systems, specially for image collections in which there is some text associated to images. Multimodality in Information Retrieval is sometimes referred to the interaction mechanisms and devices used to query the system. However, since the Multimedia Information Retrieval perspective, multimodality is referred to those methods that take advantage of different data modalities to provide access to a digital library or a multimedia collection [11, 5]. Different data modalities in multimedia are used to better understand document contents, including textual annotations, audio, images and video. In this proposal, *multimodal* will refer to the ability to represent, process and analyze two data modalities simultaneously: unstructured texts and images.

In our daily life many multimodal collections composed of unstructured text with associated images may be found, for instance, the web, which is largely composed of pages with text paragraphs and several images. Other examples of document collections with this structure are scholarly articles, book collections, news archives and medical records. The majority of these document collections are accessed nowadays using information retrieval systems devised to index text contents, so that users can search expressing their information need using textual keywords. However, the vast amount of non-text data available in many document collections may lead to the design of other effective ways for accessing and finding information.

Let's take a look to some examples of the previously mentioned multimodal document collections. Imagine you are traveling for the first time to some place and you find an interesting building that caught your attention, but you do not have any information about it. Then, you capture a photograph using your camera phone and send it to a web search service that returns a list of related web pages with historical information, technical data and tour guides [22]. This service would be useful not only for touristic places but also for products, devices and movie posters among others. Let's move to a clinical environment in which you are a physician evaluating a

patient with a medical image and, according to your experience, this image has a non-usual appearance. Then, you decide to query the medical information system in order to get a set of similar images evaluated by other physicians [13]. In addition to the similar cases, you obtain as result some recent medical papers also related to the image contents.

Both examples present situations in which the user is not able to express an accurate query using keywords. Instead, the use of the image at hand may prevent a trial-and-error loop using different keyword combinations and may offer a more precise way to access the right information. Moreover, a multimodal document collection may also be accessed using a multimodal query, i.e. a query composed of images and text descriptions. In this case, the query may lead to find highly relevant documents, since text descriptions give semantic and contextual hints about image contents and the image may help to disambiguate the right meaning of the text description [6].

The study of multimodal information retrieval systems is proposed in this document. In particular, the design of computational strategies to take advantage of multimodal interactions between image contents and unstructured text descriptions is proposed to improve the response of an image retrieval system. In addition, the evaluation of different query paradigms is proposed, including query by example, a keyword based and multimodal queries to search for images. A unified framework is proposed in this document to manage data representation, search algorithms and query resolution. The study and evaluation of kernel methods to generate Multimodal Information Spaces is proposed.

This proposal aims to approach practical and theoretical aspects of a multimodal information representation for image retrieval systems. The proposal is based on kernel methods, which provide foundations to include structure in data representation and also to combine different heterogeneous data sources. Kernel methods for pattern analysis have been studied to design machine learning algorithms, and have been widely used for non-vectorial data, such as strings, trees and graphs among others [17]. Adapting such a framework for information retrieval, and specially for multimodal information retrieval may lead to more effective systems, and also may contribute to the understanding of the relationships between information retrieval and machine learning.

2. PREVIOUS WORKS

The research field on content-based multimedia retrieval has largely grown in the past few years. Datta et al. [5] performed a simple exercise to validate this hypothesis finding a roughly exponential growth in interest in image retrieval during the last 10 years. Video and audio retrieval have attracted great interest too, leading to a more general case of information access into multimedia collections [11]. A pool of events in multimedia information retrieval have been organized to establish common test collections, evaluation protocols and baselines in a competitive environment to academically share research experiences. The following are the set of more prominent events: TRECVID [18] for content-based video retrieval, INEX [8] for structured multimedia collections, ACM-MM GrandChallenge [16] for large multimedia collections and CLEF [3] for image, video, audio and cross language collections. Importantly, most of the defined collections in these events are composed of multimodal data, and recent studies suggest the potential advantage of using multimodal synergies in image databases [5, 11].

The core strategy in Multimodal Information Retrieval is the combination or fusion of different data modalities to expand and complement information. Previously, it is important to process each independent modality. In the case of text documents, well known strategies such as the Vector Space Model are very effective and have been largely extended to solve different problems [12]. The CBIR community does not have a general agreement in the kind of image representation or retrieval model that may be applied. However, recent experimental evaluations are giving a more clear panorama of the representation problem [2], suggesting some promising directions. In fact, some of them are becoming popular in the current research such as the bag of features and statistical signatures.

Once each modality is processed to extract the most informative data, the combination procedure is applied. Two ways to combine multimodal information can be identified: late fusion and early fusion. Late fusion refers to those methods that preserve each data modality separately and, when a user request is received, two search algorithms are executed, one on each modality, to integrate the results just before deliver them to the user. On the other hand, early fusion is referred to those methods that integrate both data modalities before a user request is received, i.e. data have been previously fused and the search algorithm runs on the new fused representation. Both fusion strategies have been subject of recent research and they have provided important insights for the operation of Multimodal Information Retrieval systems.

Late fusion, i.e. combining different rankings, is also referred to as rank aggregation or data fusion [14]. In information retrieval, data fusion merges the retrieval results of multiple systems and aims at achieving a performance better than all systems involved in the process. There are several algorithms to combine rankings that are well known in the information retrieval community, such as linear combination of rankings, summing all similarity scores for each document and voting algorithms inspired in social sciences among others. These algorithms have been evaluated in text information retrieval showing an improved operation [21]. Other simple algorithms based on sets operations to merge ranking lists have been evaluated for image retrieval, using a text search engine and a content-based image retrieval system [19]. In addition, Lau et al. [9] showed that linear combinations of text and visual rankings may lead to better results than each individual system.

Early fusion aims to build an integrated representation of multimodal data to take advantage of implicit relationships. The most simple approach is to normalize and concatenate feature vector representations of each modality. Ayache et al. [1] evaluated this approach to index a video collection with multimodal information. For image retrieval this approach has also been evaluated and extended using Latent Semantic Indexing [15]. An image is considered a document with text data in a vector space model and and visual patterns represented by a bag of features. Both representations are projected together to a latent space in which the search for similar images is performed. Canonical Correlation Analysis (CCA) has also been proposed to find relationships between visual patterns and text descriptions. For instance, Vinokourov et al. [20] applied Kernel CCA to a web image collection to identify links between visual and text representations in order to solve cross modal queries. More recently, the problem of early fusion has been reformulated as a subspace learning problem that offers both dimensionality reduction and feature fusion [7]. The general problem of feature fusion is of great interest in multimedia processing for applications in classification and retrieval tasks.

3. RESEARCH PROBLEM

The main three strategies for Multimodal Information Retrieval are: semantic image retrieval, late multimodal data fusion and early multimodal data fusion. We argue that translation and auto-annotation models for semantic image retrieval lose information of image structure summarizing it into keywords. Then, visual appearance, scene composition and other visual hints are simply discarded. On the other hand, the late fusion approach uses simple strategies to combine the results such as linear combinations or voting algorithms, and the interactions between texts and images may be more complex than that.

The proposed research focuses on strategies for early multimodal data fusion to model interactions between different data modalities. A Multimodal IR system under that approach has three main associated issues as follows:

1. *Content representation of each modality.* The content representation involves the analysis and extraction of information from each modality separately. The processing of image contents and text documents is the main task in this step. It allows the filtering of non-useful data and capture the most discriminative content as is usually done in information retrieval systems.
2. *Information fusion.* The information fusion step, a particular aspect of Multimodal Information Retrieval systems, leads to the design of methods to find and represent the relationships

between both modalities. How to discover the most meaningful associations between images and text and how to complete missing data or non-clear relationships, are the main problems in this step. In this research, the design of early fusion methods is proposed, so at the end of this step, a new document representation is obtained containing both, visual and textual information.

3. *Multimodal retrieval algorithms.* Multimodal retrieval algorithms on the fused representation are designed to identify the most relevant results for the user. The main research questions in this step are related to the query representation and how to solve unimodal and multimodal queries.

4. PROPOSED RESEARCH

The main goal of the proposed research is to design and evaluate a Multimodal Information Space for content-based image retrieval. The construction of such a space is based on kernel methods, that provide a strong theoretical frame to work with different complex and structured data representations. Kernel methods have had a great impact in machine learning and pattern recognition, since they provide effective algorithms and strong theoretical properties. Shawe-Taylor & Cristianini [17] present four principles of a kernel method solution that will be followed in this proposal to approach the problem of Multimodal Information Retrieval systems:

1. *Data items are embedded into a vector space called the feature space.*
2. *Linear relations are sought among the images of the data items in the feature space.*
3. *The algorithms are implemented in such a way that the coordinates of the embedded points are not needed, only their pairwise inner products.*
4. *The pairwise inner products can be computed efficiently directly from the original data items using a kernel function.*

The following subsections present the outline of the main theoretical properties that are considered to tackle the problems of how to represent image and text document contents, how to address the fusion and combination problem using kernels and how to solve queries in a Multimodal Information Space.

4.1. Content Representation

The content representation will follow the first and fourth principles of a kernel method solution. Using the first principle we have a vector space for each data modality in the Multimodal Information Retrieval system. The key point is that the vector space is implicitly defined by the kernel function for the data that is being analyzed. A kernel function gives a similarity notion between the input data. So that, following the fourth principle, we can devise efficient methods to embed the input data into that vector space without explicitly define it. Kernel functions have attracted a lot of attention in different pattern recognition tasks, such as structure prediction in bioinformatics, text categorization and image classification. And since a kernel function provides a mechanism to introduce a similarity measure of two objects in the learning system, many of the proposed algorithms to calculate the kernel value take into account the object structure. In that way, the feature space in which the data is actually represented, is usually a high dimensional vector space that contains information about the structure and the content of the original object.

In the particular case of image and text processing, different kernel functions have been proposed. For instance, the computer vision community has developed some kernel functions to represent images using local appearance and global structure, e.g. the Spatial Pyramid Match Kernel [10]. Other kernel functions for images include histogram kernels, segmentation graph kernels and spectral based kernels among others. On the other hand, for textual documents and strings some kernel functions have also been defined to capture syntactical structure and semantic relationships [17]. Those kernel functions on both, visual and textual data, have shown effectiveness and robustness in challenging classification tasks, obtaining state-of-the-art performance. This suggest that, those kernel functions may have also a good performance in an

information retrieval task. Importantly, those functions generate a vector space to represent data of each modality.

4.2. Information Fusion

Information fusion will follow the first and second principles of a kernel method solution. According to the previous subsection, using a kernel function for text and another one for images, we have two vector spaces in which each modality is represented independently. Under a kernel framework, there are different strategies and algorithms to operate in two vector spaces simultaneously depending on the pattern of interest. So for example, if we would like to combine two vector spaces, we can generate a new vector space containing the information and structure of both modalities just by adding the corresponding kernel functions. Imagine a new vector space with structural information of images and semantic meanings of texts. This space may have million of dimensions to represent such contents, the good news is that we do not need to explicitly calculate a matching function between those vectors, instead, operate with kernel functions.

A set of operations with two kernels have been identified such that the resulting function is a valid kernel too [17]. Some of those operations include, addition, multiplication and composition among others. Depending on the operation used to combine two kernel functions, the vector space associated to the new kernel may be of a higher dimensionality. In addition, each dimension in this new vector space may be weighted in different ways. According to the fourth principle of a kernel method solution, the feature space is devised to provide linear relations among data items in the feature space. This also suggest the use of other pattern analysis algorithms to identify more meaningful relationships between images and texts in the feature space.

4.3. Information Retrieval

The Multimodal Information Retrieval algorithms will follow the third principle of a kernel method solution: algorithms only need the inner product information between vectors in the feature space, instead of the explicit vectors. In that way, we can imagine again a vector space with image and text information. If we want to calculate the matching or the similarity between a stored document and a query in the feature space, we can use the kernel functions to do so. Suppose the desired measure to be applied is cosine similarity in the Multimodal Information Space. Since a kernel function is the dot product of two elements in the feature space, we can easily calculate the cosine similarity using the following expression: $\cos(x, y) = k(x, y) / \sqrt{k(x, x)k(y, y)}$. It would directly simulate the ranking obtained in a traditional information system if our feature space is the vector space associated with term frequencies and the kernel function is the identity.

Other more sophisticated algorithms may be applied in the Multimodal Information Space, since the mathematical framework of kernel methods has found direct relationships between the Vector Space Model and a feature space. For instance, Latent Semantic Analysis may be applied in the Multimodal Information Space using Latent Semantic Kernels [4], then, we can select a subspace to project the multimodal data in which each dimension stands for a latent semantic topic in the collection. The availability of these tools show that many of the operations that are currently applicable to a Vector Space Model in Information Retrieval can be extended to a feature space, that has been called the Multimodal Information Space through this proposal.

4.4. Evaluation

There are a lot of document collections that include both, images and texts, in which users require to find information either illustrated in images or described in texts. This project has as goal to index the information of images and texts simultaneously to find relevant information independently of its original format. Although the kind of collections on which such a system may be applied is very diverse, this project aims to evaluate the proposed system in a collection of medical information, including images, medical records and scholarly papers. In particular, the collections provided in the ImageCLEFmed competition are planned to be used as well as the datasets collected in the Bioingenium Research Group, product of its operation.

A prototype system will be implemented to operate with the proposed methods, particularly to search for relevant documents given multimodal or unimodal queries. The response of the system will be assessed using standard IR measures to compare results with reported baselines and state-of-the-art methods. The response of the proposed Multimodal Information Retrieval will be also compared with the response of a standard text search engine and a standard image retrieval system to evaluate their relative performance. It is expected that the Multimodal Information Space provide more accurate results.

4.5. Performance Considerations

The proposed research is mainly based on kernel methods that may work on very high dimensional spaces. Kernel based algorithms do not need to operate explicitly in the high dimensional space, and that leads to the implementation of fast similarity measures between structured data. For example, the Pyramid Match Kernel [10], used to approximate the matching between two sets of image features, provides high accuracy and low computational effort compared to the optimal correspondences between the sets' features.

However some learning algorithms need to process a kernel matrix that grows quadratically with the size of the sample. For instance, a Singular Value Decomposition (SVD) of the kernel matrix is useful for doing principal component analysis or latent semantic analysis [17]. But the SVD algorithm is $O(n^3)$ and it would demand huge computational resources or may take a long time to process for large data collections. The complexity of the proposed algorithms will be studied to evaluate the impact on the system performance. The majority of the algorithms that require to process a kernel matrix are training algorithms that can be executed offline. Moreover, training algorithms are not needed to be applied on the complete document collection. That is, a representative sample may be taken from the collection to analyze patterns, structure and relationships, and later the obtained models may be generalized to the whole collection. When possible, parallel or distributed implementations will be considered for algorithms with high complexity.

5. SUMMARY

This paper has presented a research agenda to study and evaluate Multimodal Information Spaces for Content-Based Image Retrieval. The main research question is how can we retrieve visual information from a large multimodal document collection, taking into account that both visual and textual contents may provide useful information to improve the retrieval performance. The use of kernel functions to construct Multimodal Information Spaces is proposed, and a framework based on kernel method solutions will be followed.

Under the proposed framework, different image and text features may be fused in a high-dimensional space, in which a search algorithm may be designed. Each data modality in an image collection will be processed independently and will be integrated using the proposed framework. The image collection to be used is taken from the medical domain in which the multimodal structure may be found in health records and scholarly articles. The evaluation and analysis of standard information retrieval measures is also proposed to assess the contribution of the proposed research.

Bibliography

- [1] Stéphane Ayache, Georges Quénot, and Jérôme Gensel. Classifier fusion for svm-based multimedia semantic indexing. pages 494–504. 2007.
- [2] A. Bosch, X. Munoz, and R. Martí. Which is the best way to organize/classify images by content? *Image and vision computing*, 25(6):778–791, 2007.
- [3] Martin Braschler and Carol Peters. Cross-language evaluation forum: Objectives, results, achievements. *Information Retrieval*, 7(1):7–31, January 2004.
- [4] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. Latent semantic kernels. *Journal of Intelligent Information Systems*, 18(2):127–152, March 2002.

- [5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, April 2008.
- [6] Xin Fan, Xing Xie, Zhiwei Li, Mingjing Li, and Wei-Ying Ma. Photo-to-search: using multimodal queries to search the web from mobile devices. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 143–150, Hilton, Singapore, 2005. ACM.
- [7] Yun Fu, Liangliang Cao, Guodong Guo, and Thomas S. Huang. Multiple feature fusion by subspace learning. In *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 127–134, New York, NY, USA, 2008. ACM.
- [8] Norbert Gövert and Gabriella Kazai. Overview of the initiative for the evaluation of xml retrieval (inex) 2002. In *In Fuhr et al*, volume 6, pages 1–17, 2003.
- [9] C. Lau, D. Tjondronegoro, J. Zhang, S. Geva, and Y. Liu. Fusing visual and textual retrieval techniques to effectively search large collections of wikipedia images. pages 345–357. 2007.
- [10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society.
- [11] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, February 2006.
- [12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [13] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, February 2004.
- [14] Rabia Nuray and Fazli Can. Automatic ranking of information retrieval systems using data fusion. *Inf. Process. Manage.*, 42(3):595–614, May 2006.
- [15] Trong-Ton Pham, Nicolas Maillot, Joo-Hwee Lim, and Jean-Pierre Chevallet. Latent semantic fusion model for image retrieval and annotation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 444, 439. ACM, 2007.
- [16] Lawrence A. Rowe and Ramesh Jain. Acm sigmm retreat report on future directions in multimedia research. *ACM Trans. Multimedia Comput. Commun. Appl.*, 1(1):3–13, February 2005.
- [17] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.
- [18] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, NY, USA, 2006. ACM.
- [19] Julio Villena-Román, Sara Lana-Serrano, and José C. González-Cristóba. Miracle at imageclefmed 2007: Merging textual and visual strategies to improve medical image retrieval. 2007.
- [20] Alexei Vinokourov, David R. Hardoon, and John Shawe-Taylor. Learning the semantics of multimedia content with application to web image retrieval and classification. In *In Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, 2003.
- [21] Shengli Wu. Applying statistical principles to data fusion in information retrieval. *Expert Systems with Applications*, In Press, Corrected Proof, 2008.
- [22] Tom Yeh, Kristen Grauman, Konrad Tollmar, and Trevor Darrell. A picture is worth a thousand keywords: image-based object search on a mobile platform. In *CHI '05 extended abstracts on Human factors in computing systems*, pages 2025–2028, Portland, OR, USA, 2005. ACM.