COMBINING VISUAL FEATURES AND TEXT DATA FOR MEDICAL IMAGE RETRIEVAL USING LATENT SEMANTIC KERNELS

Juan C. Caicedo Jose G. Moreno Edwin A. Niño Fabio A. González

ABSTRACT

In this paper we propose an strategy to fuse visual features and unstructured-text data in a medical image retrieval system. The main goal of this work is to investigate whether the semantic information from text descriptions can be transferred to a visual similarity measure. Then, a system to search using the query-by-example paradigm is evaluated instead of a keyword-based search. We achieve this by using Latent Semantic Kernels to generate a new representation space whose coordinates define latent concepts that merge visual patterns and textual terms. The proposed method is tested in a medical image collection from the ImageCLEFmed08 challenge. The experimental evaluation tests the system using different image queries. The results show an improvement of the visual-text fused approach with respect to only using visual information. Keywords— Image Retrieval, Latent Semantic Kernels

1. INTRODUCTION

Large amounts of digital images are constantly acquired in different contexts of our daily life, from personal photos taken with digital cameras and mobile phones, to high resolution images used in advertising and journalism. The design of systems and models to provide access to a large number of images has been an active research area during the last years. Image retrieval using visual content is still an open problem due to the semantic gap [16], i.e. the disagreement between visual features and conceptual interpretations. To overcome this problem, the content-based image retrieval (CBIR) community has proposed different strategies for learning to understand visual image content and automatically annotate images using a wide variety of concepts [5, 8]. A prerequisite for automatic image annotation is the presence of reliable and explicit image metadata that clearly explains image content [5]. These annotations may be given by human beings that observe each image and describe its content in terms of an ontology or predefined dictionary. However, this process is very expensive and even infeasible for very large collections.

The main purpose of collecting human annotations is to train a model that learns the association between visual patterns and semantic descriptions [2, 3]. Several image collections may be extracted from unstructured document collections, therefore, each image has a natural language textual description associated to it. Examples of those collections are images in books, scientific papers, or web pages. During the writing process, people select the pictures to be placed into the documents and reference them inside the text, explaining their meaning or context. This process is performed by the majority of the writers either if the document is a web page or a scholarly article. Both images and text bring complete meaning to the document and they are complementary information for the reader.

Image retrieval systems can exploit text descriptions to search only using keywords. Under this approach, text-based image retrieval may be a good strategy to retrieve related images when the user knows an appropriate keyword combination to express his information need. Most of the current commercial image retrieval systems are based on this strategy. However, sometimes they fail to retrieve relevant images, since their indexing method only relies on text annotations, and the image visual content is completely ignored. On the other hand, despite of the existence of text annotations, 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. For instance, image retrieval systems only based on text annotations cannot support a query-by-example paradigm, in which a user presents an example image to retrieve related pictures from the database [14].

There are several scenarios in which query-by-example may be useful for image and information retrieval. Imagine that a user is traveling for the first time to some place and finds an interesting building that catches his attention, but he does not have any information about it. Then, the user captures a photograph using his camera phone and send it to a web search service that returns a list of related images with their attached text [17]. A similar event is experienced in a clinical environment in which the user is a physician evaluating a patient with a medical image and, according to his experience, this image has a non-usual appearance. Then, the physician decides to query the medical information system in order to get a set of similar images evaluated by other physicians [12]. The physician obtains as result some clinical records with similar images and also recent medical papers related to the image content. Both examples present situ-

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.

ations 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.

In this work, we do not consider a system to retrieve images using keyword-based queries, instead, the use of a query-by-example paradigm is studied. An image retrieval system based on the query-by-example paradigm cannot directly use the text annotations attached to images in the database, since users do not provide any keyword. If the system do not have an automatic image annotation module to generate keywords for the query image, the natural way to go is to use only visual similarities to identify relevant results. However, the semantic gap would still be there, even though there is a large amount of text related to images in the database. We want to investigate whether the information of text documents may be transferred to a content-based image retrieval system in order to enrich the visual information representation.

We propose a strategy based on Latent Semantic Kernels (LSK) to approach the image-text fusion problem, using kernel functions on visual and textual contents. Under this framework, images and texts may be represented by structured data, since kernel functions allow to embed those objects into a high-dimensional feature space. Both feature spaces are combined in a unique representation space applying operations on the kernel functions. Then, an Eigendecomposition of the data in the new high-dimensional space is calculated to identify latent semantic correlations between textual and visual features. Using this information, a latent semantic space is defined in which textual and visual features are projected together. In that way, text semantics is included in the visual-based search engine to improve the system response, when compared to a system that only uses visual features.

Since the queries that are considered in this work do not have text information but are composed of example images, visual features of each query are projected into the latent semantic space, in such a way that they automatically correlate with text data in the fused representation. Then, the most similar images in that space are retrieved. The proposed system was evaluated using a medical image database extracted from biomedical articles that was used in the ImageCLEFmed 2008 challenge to evaluate the performance of image retrieval systems in the medical domain. Our approach, for searching with image examples, shows that unstructured-text information can be effectively included in a visual retrieval system to enhance the image representation. This paper is organized as follows: Section 2 presents a brief review of the related work. Section 3 describes the methods and models applied in the proposed strategy. The experimental setup is presented in Section 4. Section 5 shows the results of the experimental evaluation and finally, Section 6 presents some concluding remarks.

2. RELATED WORK

Approaches to automatically annotate images have been actively researched during the recent years. The main purpose of these approaches is to assign a set of keywords to each image, trying to describe, as accurately as possible, the content of each image. Different strategies such as co-occurrence models [10] and machine translation models [2] between image sub-blocks and text terms have been proposed. Also relevance models have been used to expand the probability of visual patterns given a particular set of keywords to search images [6]. The kind of queries that they support are of the form "find all the images of tigers in grass".

The main assumption of these approaches is that the system will operate on the basis of queries with keywords. In a similar way as cross-lingual text information retrieval, these models are often known as cross-media strategies, in which a user provides a set of keywords and a set of images is retrieved. However, under this scheme image representation is usually reduced to keywords to apply standard text retrieval approaches.

In information retrieval, combining information from different data sources is known as data fusion, that is, the use of complementary information to solve queries. Two main strategies are considered in data fusion according to the time in which the combination is achieved: late fusion and early fusion. Late fusion deals with the combination of rankings from two different search engines, after each one has processed the corresponding part of the query. Early fusion deals with the construction of a new document representation that contains both types of information simultaneously. The proposed approach in this paper is related to early fusion.

Latent semantic analysis has been used by Pham et al. [13] for early fusion of image features and keywords. They processed images using a bag-of-features approach in which each image is represented by the histogram of occurrences of local visual patterns in a dictionary. This representation is concatenated with a vector space model for text data, and correlations between visual patterns and keywords are obtained applying latent semantic indexing. Their results showed a promising application of this approach for image retrieval. However, this evaluation has been conducted using text queries or multimodal queries, composed of both images and keywords. On the other hand, the collection has some category labels attached to each image so the vocabulary has been restricted and controlled to express several concepts on the image collection. In the present work, we attempt to approach the problem of combining unstructured text with visual features in a unified search index.

The evaluation followed in this paper is based on a collection of medical images extracted from biomedical articles, that constitutes the dataset used on the ImageCLEFmed 2008 challenge. This dataset contains more than 67,000 images with their associated captions extracted from papers. During the challenge, a set of 30 topics are proposed by the organizers, whose baseline results have been evaluated by several assessors. The 30 topics are divided in three categories: visual topics, mixed topics and semantic topics. When participants got involved in the challenge, they are asked to solve all the 30 topics, indicating whether the applied methods require only visual data, only text data or both. In the category in which researchers used only visual data to query the system, the best method reached a Mean Average Precision (MAP) of 0.04 that year. This suggest that the retrieval task is particularly challenging when the system uses only visual data. This is the category in which our results are contextualized since we use only visual data to query the system.

3. LATENT SEMANTIC FUSION MODEL

Latent Semantic Indexing (LSI) has been proposed in information retrieval to model the statistical correlation between terms. It has been applied to find hidden semantic concepts written with different synonyms and also for cross-lingual information retrieval to search documents in another language. LSI follows the same principle of a generalized vector space model, with a special choice of the co-occurrence matrix, given by the eigen-decomposition of the term-document matrix. Conceptually, the implementation of LSI requires a term frequency representation of each document in a vector space model, so that the method can identify highly correlated dimensions on the training dataset. Therefore, images must be modeled as feature vectors to be able to correlate visual features and text terms. Both feature vectors are concatenated in the same term-document matrix, and the co-occurrence matrix can be directly calculated. Following this approach, limits the kind of data structures that may be used to represent image contents, such as histograms, trees, sets of points or graphs, that would not be allowed since the method is restricted to explicit vector representations only.

The use of structured image representations has been shown to offer state-of-the-art performance in different computer vision tasks, for instance in natural image categorization problems in which spatial clues are taken into account to identify the correct category [1]. Interestingly, some structured image representations can be embedded in high-dimensional feature spaces using the kernel trick from the machine learning literature, provided that an appropriate similarity function is devised.

We propose the use of kernel functions instead of vector descriptors to represent both, image and text features, to address the information fusion problem. Intuitively, kernel functions are similarity measures between data objects that implicitly define a high-dimensional feature space to represent the input objects. This property enables the proposed solution to be independent of the particular data type or data structure used to represent the information. Different learning algorithms have been formulated to work with kernel functions instead of explicit vector representations, such as Support Vector Machines, kernel-based clustering algorithms and others. In particular, Cristianini et al. [4] formulated a kernel-based Latent Semantic Analysis algorithm to find the directions of maximum variance in a dataset described by a kernel function. In this Section we describe the properties of Latent Semantic Kernels (LSK) and the details of the proposed approach to combine information.

3.1 Latent Semantic Kernels

Let $\phi(d_i)$ be the representation of the document d_i in an n-dimensional vector space, $\mathbb{R}^n$, and D be the term-document matrix, an $n \times l$ matrix whose columns are the vector representations of each document $\phi(d_i)$ for $i \in 1...l$, with l being the size of the training dataset. LSI stands on a Singular Value Decomposition (SVD) of the term-document matrix as follows:

$$D' = U\Sigma V' \tag{1}$$

where U and V are orthonormal matrices containing the eigenvectors of $D'D$ and DD'respectively, and Σ is a diagonal matrix with the eigenvalues of D. Vectors in the matrix U are the basis to span the latent semantic space so that each document is projected to the latent space using:

$$d \rightarrow \phi(d)U_k \tag{2}$$

where U_k denotes a matrix with the first k eigenvectors of U according to a decreasing order given by the Eigenvalues in Σ. Hence, the parameter k establishes the dimensionality of the latent semantic space. SVD merges highly correlated terms in the same Eigenvectors, which correspond to the factors of the latent semantic space. This factors can be interpreted as concepts describing the data.

Note that V contains the eigenvectors of the matrix $DD' = K$, where K is a Gram matrix or kernel matrix. Hence, in the case in which a vector representation of the data is not available, we can consider using the Eigendecomposition of the kernel matrix, $K = V\Lambda V'$. It is clear that the needed vectors to span the latent semantic space are not available using this representation. However, it has been shown that the vectors in U can be expressed in terms of the vectors in V as a consequence of the eigendecomposition of the term-document matrix [15]. A dual representation of the projected objects in the latent semantic space can be expressed as:

$$\phi(d)U_k = \left(\lambda_i^{-\frac{1}{2}} \sum_{j=1}^{l} (v_i)_j k_m(d_j, d) \right)_{i=1}^{k} \tag{3}$$

where λ_i, v_i are the Eigenvalue, Eigenvector pairs of the kernel matrix, and $k_m(d_j, d)$ is the kernel function calculated between the j-th training document, d_j, and the new document d. Note that this dual representation does not require an explicit processing of document vectors, instead, an appropriate kernel function, defined among the input objects, is needed. Now that we have an explicit vector representation in a low-dimensional feature space for any document in the collection, the similarity measure between two projected objects can be calculated as the dot product in the latent semantic space:

$$\hat{k}(d_1, d_2) = \phi(d_1)U_k U_k' \phi(d_2) \tag{4}$$

Once the singular value decomposition of the kernel matrix has been calculated, the new representation $\phi(d)U_k$ can be efficiently computed for all the documents in the collection, and also for any new document or query coming into the system. The transformation matrix $V\Lambda^{-1/2}$ can be precomputed and applied to a vector of kernel values between the current document and the training items.

3.2 Fusing visual and textual data

A document d, in this context, is defined as a data structure that contains an image and some textual annotations. Let $k_v(d_1, d_2)$ be a kernel function that evaluates a visual similarity measure between the images in the first document, d_1, and the second document, d_2. Let $k_t(d_1, d_2)$ be another kernel function that evaluates a similarity measure between the text annotations in the document d_1 and the document d_2. These two kernel functions allow to project the contents of documents into two different high-dimensional feature spaces, one for visual data and one for text data. The first step to combine the information in both feature spaces is to define a new kernel function as a combination of the two available ones:

$$k_m(d_1, d_2) = k_v(d_1, d_2) + k_t(d_1, d_2) \quad (5)$$

The feature space induced by this kernel function is a high-dimensional vector space in which the coordinates of the first space are concatenated with the coordinates of the second space. Note that more complex combination functions may be designed to combine the information between visual features and text data. A simple additional operation on the new kernel function may be the composition of this function with a polynomial construction, in which the interactions of the terms in the text representation space and the features in the visual representation space are taken into account.

Once the combined feature space has been defined, a training dataset is selected to identify the correlations between both document views, i.e. between visual features and text data, so the kernel matrix K with the values of the combined kernel function must be computed for this set. The visual features and text terms that co-occur frequently will tend to align with the same eigenvectors, since LSK identifies the directions of maximal variance in the dataset. In the same way as term correlations emerge from the dataset when using LSI, in this case, LSK finds implicit correlations between visual features and text terms.

3.3 Solving queries

In this work, the use of the query-by-example paradigm is proposed to search images. Then, when a user sends a query, the system can only extract visual features from it. The proposed approach is intended to automatically correlate visual features with text terms in the database, even though the user does not provide any keyword that describes the image content. Let q be the query document, in this case, an image example. Since queries are expressed only in terms of the visual data, the combined kernel in Equation 5 is reduced to:

$$k_m(q, d_i) = k_v(q, d_i) \quad (6)$$

with d_i a document in the training set. The reduced form of the combined kernel is used to project the image query into the latent semantic space, according to Equation 3. That is, once the query is represented in the latent semantic space, their content is now expressed using a set of features that account the correlation between both data types. Images in the database have been projected using the same strategy, but using text annotations when they are available. In that sense, the proposed strategy helps to include text semantics in the image search index. The final ranking of the database documents with respect to the query document is calculated according to Equation 4, that is, the dot product in the latent semantic space.

3.4 Performance considerations

The proposed strategy is based on the Eigendecomposition of a kernel matrix. The computational cost of this decomposition is significant and it is not feasible to calculate it on a very large document collection. To make this strategy usable, two considerations have to be taken into account. First, a sampling strategy may be applied on the document collection to obtain a representative sample as training set. When the Eigenvectors are obtained, they can be used to project the remaining part of the document collection as well as new documents included in the future. Second, this algorithm is applied off-line and only once to index the complete document collection. Naturally, a new calculation over a more appropriate sample may be calculated when needed, to re-index the complete collection.

Note that once the Eigenvectors have been calculated, a transformation matrix can be precomputed to project each document to the latent semantic space, as claimed at the end of the Subsection 3.1. Then, the projection operation requires the computation of the kernel function between the current document and the training dataset, before the transformation matrix can be applied. This operation is linear in the size of the training set.

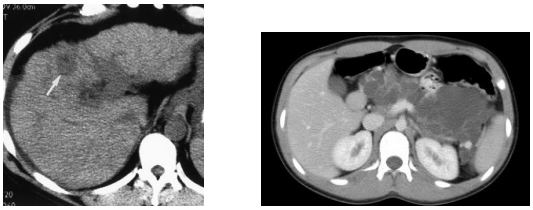

Figure 1: Two relevant images for the query topic: *Show me abdominal CT images showing liver blood vessels*

4. EXPERIMENTAL SETUP

4.1 Dataset

We evaluated the proposed method using as dataset the image collection from the ImageCLEFmed 2008 challenge. This dataset is composed of 67,115 medical images from different modalities, including x-rays, magnetic resonance images, dermatology and microscopy images, among others. Those images have been obtained from a database of biomedical papers, and each image in the dataset is distributed with an XML file with the caption extracted from the article. Together with the image collection, the organizers of this competition provide a set of 30 topics composed of several images and a text description. Each topic is composed of 1, 2 or 3 example images, that accounts to a total of 61 different query images. In addition, the relevant judgments for each topic have been provided by the competition organizers and have been used in this evaluation to calculate the performance of the proposed strategies. Figure 1 shows some of the query topics followed by several relevant results according to the relevant judgements.

The ImageCLEFmed challenge evaluates the performance of the proposed approach according to the underlying strategy to solve the given topics. In general, there are three different approaches to solve queries in this evaluation: first, using only visual information extracted from example images; second, using only textual information; and third, using both, the image and the text annotation on the query. We are interested in evaluating the retrieval performance when only a visual query is provided. In that sense, our approach to search images falls on the category of only visual approaches, since the text in the query is not provided to the search algorithm. The best performance in such a category in the 2008 challenge was of 0.04 in the MAP measure.

As mentioned before, the suggested queries for this dataset

have several image examples each. We investigate the performance of the proposed strategy using all suggested images per query, as well as the performance using only one image at a time. This may be understood as modelling different user behaviours in which sometimes users have different images for the same query, such as in an electronic health record, that might be useful to find appropriate results. On the other hand, queries with individual images may be understood as concrete needs to retrieve similar images and to realize visual patterns with respect to other images.

4.2 Visual and Textual kernels

We implemented the two required kernel functions k_v for visual data and k_t for text data. First, to illustrate the advantage of using structured image representations, we evaluated the performance of two image descriptors. First, a simple vector descriptor has been used to encode global image characteristics using a downscaled image representation. The image is re-sized to 32×32 pixels independently of the original aspect ratio. The color information of each pixel is preserved in the RGB color space. Then, the feature vector is build using $32 \times 32 \times 3$ features. This representation may be useful to preserve some information about spatial image relationships. The kernel function applied to this image descriptor is the cosine kernel, defined as:

$$k_v(i_1, i_2) = \frac{< i_1, i_2 >}{|i_1||i_2|} \quad (7)$$

The second image representation is an spatial extension of the bag-of-features approach following the model of Lazebnik et al. [7]. Local patches of 16×16 pixels are extracted from images using a regular grid with an offset of 8 pixels. From each patch, the SIFT [9] descriptor is computed. Then, a sample of patches is drawn from the training set to build a dictionary of visual patterns using the k-means algorithm. The image representation is organized on 3 partition levels, following a pyramid structure, in which the base is the complete image, and the subsequent levels split each region in the previous level in four equal-sized new regions. For each region in the pyramid a histogram accounting the occurrence of each visual pattern in the dictionary is constructed. Then, the Spatial Pyramid Kernel is computed as:

$$k_v(i_1, i_2) = \sum_{l=0}^{L} \sum_{j=1}^{2^{2*l}} \sum_{m=1}^{D_l} min(H_{i_1}^j(m), H_{i_2}^j(m)) \quad (8)$$

Notice that one important parameter of this kernel is the size of the visual patterns dictionary. We experimentally evaluated different dictionary sizes to determine a good retrieval performance. Text captions for each image are processed as individual documents. A typical information retrieval pre-processing is applied on the text collection, including stop-words removal and stemming. Then, a vector space model is built to index the frequency of document terms. We applied TF-IDF weighting to the collection, and finally, the cosine kernel is computed in the same way as explained in Equation 7.

4.3 Kernel Combination

From the collection of 67,115 images, a subset of 20,000 images has been randomly chosen as training dataset. Using the training set, the kernel matrix is constructed to compute its eigendecomposition. The combined kernel function is a linear combination of the visual kernel k_v and the textual kernel k_t with equal weights as was presented in 3.2. This is called the linear kernel construction in the following Subsections. This kernel function has also been composed with a polynomial and Gaussian constructions. The polynomial construction is obtained as follows:

$$k_p(d_1, d_2) = (k(d_1, d_2) + 1)^p \quad (9)$$

where p is the degree of the polynomial and has been set to two in our experiments. This construction is equivalent to span a feature space in which the interactions between pairs of features are taken into account. That is, the polynomial construction indexes the interactions between each visual feature and each text term. The Gaussian construction is obtained as follows:

$$k_g(d_1, d_2) = exp\left(\frac{k(d_1, d_1) + k(d_2, d_2) - 2k(d_1, d_2)}{\sigma^2}\right) \quad (10)$$

where σ is the parameter of the function. Experiments using $\sigma^2 \in 0.1, 1, 10$ were run. With each kernel construction, the eigenvalues and eigenvectors were calculated on the training dataset. Then, the collection was indexed using the projection operation selecting different sizes of the latent semantic space. The dimensionality of the latent semantic space was evaluated using powers of two, starting with 2^0 to 2^{14}, to analyze the generalization ability of the proposed strategy. To rank images from the database, the dot product in the latent semantic space is used.

4.4 Performance Evaluation

We used a variety of performance measures to evaluate the response of the proposed image retrieval system. These metrics include the Mean Average Precision (MAP) and the Recall value in the first 1,000 results. Also, we examined the behaviour of the system using P10, P20, P50 and P100. R-recall and number of relevant images retrieved is also reported for comparison of different experiments. We compare different configurations of the proposed approach with respect to the performance of visual similarity measures without any other processing, since this would be the approach to search when only visual queries are received in the system.

5. RESULTS

5.1 Image Representation

The problem of including discriminative visual information in a CBIR system has been widely investigated. It has been also realized that more visual information does not necessarily lead to more semantic results. The goal of our first experiments was to determine the impact of two different strategies to search directly using only visual similarity, as well as evaluating the performance of the proposed strategy using different image kernels. Table 1 summarizes the main results of these experiments.

Direct search means that the kernel has been used to rank images from the database given 61 different queries. In terms of MAP and Recall the reader can notice that the performance of both kernels is similar, due mainly to the semantic gap, i.e. visual information alone does not provide enough

Representation	Direct Search MAP	LSK MAP	Direct Search Recall	LSK Recall
Color Feature Vector	0.0152	0.0159	0.1147	0.1247
Spatial Pyramid	0.0162	0.0204	0.1098	0.1386

Table 1: MAP and Recall values obtained by each image representation

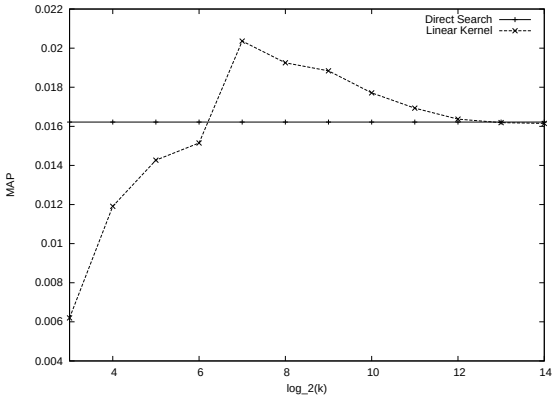

Figure 2: Evaluation of different latent space dimensions using the linear kernel construction

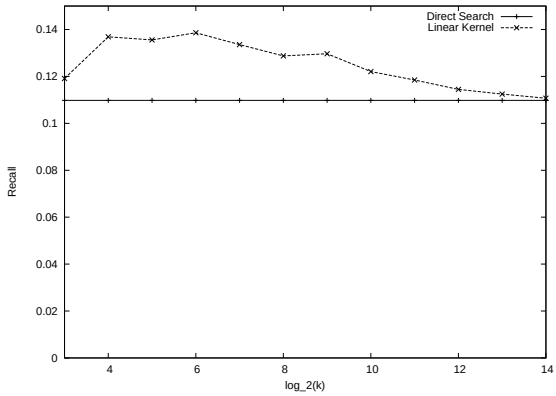

Figure 3: Evaluation of different latent space dimensions using the linear kernel construction

hints to recognize semantic content in images. The proposed LSK approach is a learning strategy to correlate text information with visual information and in this case, the performance difference is larger. The MAP value obtained by the color feature vector is almost the same when using direct matching or LSK, while the MAP value of spatial pyramid kernel improves in about 25%. This supports the idea that spatial pyramid has more discriminative information for learning purposes than the simple color feature vector, even though they have similar behaviours during direct matching. Besides, the discriminative information of the spatial pyramid comes from the particular structured data that might not be included in a latent semantic analysis without the help of kernel algorithms.

5.2 Size of the Latent Semantic Space

The size k of the latent semantic space is determined by the number of Eigenvectors involved in the projection process. The smaller the value of k, the higher the correlation between terms. On the other hand, as long as the value of k approaches the size of the training dataset which is the maximum size allowed, the latent representation is more similar to the original image representation. Figures 2 and 3 show the performance of the system in terms of MAP and Recall respectively, when the size of the latent semantic space is changed. We evaluated increasing values of k as powers of 2, starting from very small ones $2^3 = 8$ to very large ones $2^{14} = 16384$.

Figures 2and 3 show how the performance increases with k and then converges to the baseline performance using only visual information. When the size of the latent semantic space is appropriate enough, the correlation between features and text is meaningful to improve the results. As long as the k value reaches the maximum allowed value, no correlation between features and text is incorporated in the representation following a similar behaviour as the one observed when only the visual information is used. These plots were obtained using the Pyramid Kernel combined with the text kernel using a linear combination.

5.3 Kernel construction

We can take advantage of different kernel constructions to induce more complex feature spaces in which text information and visual features are being correlated. We used a Polynomial construction and a Gaussian composition as was discussed in Subsection 4.3. The underlying kernel, previous to the composition, is the mixed visual-text kernel using a linear combination of the original kernel functions.

Table 2 shows the performance measures for different kernels used to compute the latent semantic space. Different aspects of the linear construction have been discussed on previous Subsections, and here other performance measures are reported together with different kernel compositions. In general, the results show that LSK improves the performance with respect to the direct image matching. However, the polynomial construction (using a polynomial degree equal 2) does not improve at all on the linear construction. The σ parameter of the Gaussian kernel was experimentally determined using different parameters of σ. This composition slightly improves the performance with respect to the linear one, but in general they are very similar. This may indicate that the source kernels are non-linear enough to find correlations in this particular problem and no further compositions are required to improve the performance.

5.4 Queries: Single Images vs. Multiple Images

We also investigated whether using multiple images to

Measure	Direct Search	Linear LSK		Polynomial		Gaussian	
MAP	0.0162	0.0204	25.6%	0.0195	20.0%	0.0205	26.5%
Precision10	0.0541	0.0590	9.1%	0.0557	3.0%	0.0607	12.1%
Precision100	0.0303	0.0392	29.2%	0.0397	30.8%	0.0393	29.7%
Recall	0.1098	0.1386	26.3%	0.1394	27.0%	0.1404	27.9%

Table 2: Best MAP, P10, P100 and Recall values obtained by each of the evaluated kernel constructions

query the system may help to identify more relevant results. The ImageCLEFmed challenge provides one or several images for each of the 30 proposed topics, leading to 61 different query images. Results in all previous Subsections have been reported using each of the 61 images as independent queries. This allowed us to evaluate if the contribution of each image for solving the proposed topics could be improved. In this Subsection, we present an evaluation of combining the results of each query image to obtain a unique result list for each of the 30 topics. Table 3 shows a summary of these results using different performance measures. The combination of search results when several query images are available follows a MAX strategy, that consists in selecting as score for each image in the dataset the maximum similarity among the different query images.

This evaluation has been done for completeness, in order to compare our results with those in the CLEF 2008 competition. We compare the MAX combination strategy with the best ranking using individual images for each topic. Selecting the best individual image is very similar to combining the results using ranking combination, in terms of MAP. However, in terms of early-precision and recall, the best individual ranking shows in general a better response. It suggest that to exploit the availability of several relevant examples in the query, algorithms able decide the image relevance with more effective rules should be used.

In both cases, the proposed LSK strategy is still able to offer a performance improvement over the direct image search. Our best general result in terms of MAP to solve the 30 suggested topics is 0.025 after mixing visual and text features. It contrasts with 0.04 [11] that was the best result using visual information, including about 80,000 features of color, texture and edges. We believe that our visual similarity measure is not good enough to reach the already established baseline, but still, despite the limited discrimination power in the proposed similarity measure, the proposed LSK framework was able to improve the results. Further experimentation using as baseline a better visual representation is needed. Also, the application of the proposed framework in other datasets will help to better understand the potential of this approach.

6. CONCLUSIONS

Fusing visual features and text terms is a very interesting approach to automatically annotate image contents for information retrieval applications, specially, when text annotations attached to images are not structured or organized in a controlled vocabulary. Unstructured text descriptions have been naturally attached to images by writers when they prepare documents for digital libraries, academic publications and web pages. This information may be exploited to improve the response of information retrieval systems. However, solving this problem is a very difficult task, since the explicit relationships between image contents and text descriptions are not explicit.

We have presented an approximation to the problem of correlating free-text with visual features for solving queries based on visual examples. This novel approach is based on a kernel method solution that allows to model complex document representations by operating with appropriate similarity measures. In particular, the Latent Semantic Kernel method has been applied on a combined representation of image descriptors and text data. This method is a dual representation of the well known Latent Semantic Indexing approach for information retrieval, which is a widely used method for content indexing.

We showed that the use of more discriminative image features may be included in this framework, even though the explicit feature space is not available to perform latent semantic analysis. This is particularly useful when dealing with complex information such as images, audio or video, since the information may be represented in a more natural way using appropriate kernel functions.

The experimental evaluation in this paper has been conducted on a large collection of medical images extracted from biomedical articles. Each image in this collection has a set of text annotations extracted from the image caption in the paper, which is used as semantic or contextual description. Althouth the evaluation was applied on medical images, the proposed strategy may also be applied to different kind of pictures. The experimental results showed that the proposed approach is able to improve the retrieval results in terms of precision and recall.

7. REFERENCES

[1] *Context-based vision system for place and object recognition* (2003).

[2] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. Matching words and pictures. *The Journal of Machine Learning Research 3* (2003), 1107–1135.

[3] Cheng, S. F., Chen, W., and Sundaram, H. Semantic visual templates: linking visual features to semantics. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on* (1998), pp. 531–535 vol.3.

[4] Cristianini, N., Shawe-Taylor, J., and Lodhi, H. Latent semantic kernels. *Journal of Intelligent Information Systems 18*, 2 (March 2002), 127–152.

[5] Datta, R., Joshi, D., Li, J., and Wang, J. Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv. 40*, 2 (April 2008), 1–60.

[6] Jeon, J., Lavrenko, V., and Manmatha, R. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in*

Measure	MAX Direct Search	MAX Linear LSK		Best Individual Ranking Direct Search	Best Individual Ranking Linear LSK	
MAP	0.0225	0.0251	11.1%	0.0254	0.0281	10.5%
Precision10	0.0500	0.0567	13.3%	0.0800	0.0933	16.7%
Precision100	0.0293	0.0360	22.7%	0.0473	0.0570	20.4%
Recall	0.1423	0.1434	0.9%	0.1624	0.1795	10.5%

Table 3: Evaluation of different performance measures for combining the results.

informaion retrieval (New York, NY, USA, 2003), ACM, pp. 119–126.

[7] Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (2006), vol. 2, pp. 2169–2178.

[8] Liu, Y., Zhang, D., Lu, G., and Ma, W. Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition 40*, 1 (2007), 262–282.

[9] Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*, 2 (November 2004), 91–110.

[10] Mori, Y., Takahashi, H., and Oka, R. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management* (1999).

[11] Müller, H., Kalpathy-Cramer, J., Kahn, C., Hatt, W., Bedrick, S., and Hersh, W. Overview of the imageclefmed 2008 medical image retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access*, C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas, and V. Petras, Eds., vol. 5706. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, ch. 63, pp. 512–522.

[12] Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics 73*, 1 (February 2004), 1–23.

[13] Pham, T., Maillot, N., Lim, J., and Chevallet, J. 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* (2007), ACM, pp. 444, 439.

[14] Rasiwasia, N., Moreno, P., and Vasconcelos, N. Bridging the gap: Query by semantic example. *Multimedia, IEEE Transactions on 9*, 5 (Aug. 2007), 923–938.

[15] Shawe-Taylor, J., and Cristianini, N. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[16] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence 22*, 12 (2000), 1349–1380.

[17] Yeh, T., Grauman, K., Tollmar, K., and Darrell, T. 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* (Portland, OR, USA, 2005), ACM, pp. 2025–2028.