Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization

Juan C. Caicedo\textsuperscript{a}, Jaafar BenAbdallah\textsuperscript{b}, Fabio A. González\textsuperscript{a,∗}, Olfa Nasraoui\textsuperscript{b}

\textsuperscript{a}Computer Systems and Industrial Engineering Department, National University of Colombia, Cra 30 45 - 03, Ciudad Universitaria, Edif. 453, Of. 114. Bogotá, Colombia

\textsuperscript{b}Department of Computer Engineering and Computer Science, University of Louisville, Louisville KY, USA

\begin{abstract}
Massive image collections are increasingly available on the Web. These collections often incorporate complementary non-visual data such as text descriptions, comments, user ratings and tags. These additional data modalities may provide a semantic complement to the image visual content, which could improve the performance of different image content analysis tasks. This paper presents a novel method based on non-negative matrix factorization to generate multimodal image representations that integrate visual features and text information. The proposed approach discovers a set of latent factors that correlate multimodal data in the same representation space. We evaluated the potential of this multimodal image representation in various tasks associated to image indexing and search. Experimental results show that the proposed method highly outperforms the response of the system in both tasks, when compared to multimodal latent semantic spaces generated by a singular value decomposition.
\end{abstract}

1. Introduction

Most of the effort on web-content mining has concentrated on textual (and hyper-textual) data. However, visual information is an important component of the web content nowadays. In particular, the advent of the Web 2.0 has been accompanied by an explosion of multimedia content. Specialized sites, such as Flickr and Picassa, host billions of pictures uploaded by users. Other types of sites that allow users to upload visual content include: social networking sites, such as Facebook and MySpace, community-generated content, such as Wikipedia, and individual-generated content, such as in the blogosphere and Twitter.

The most salient characteristic of web image collections is that they come with a wide variety of associated data, such as text descriptions, tags, ratings and user comments. The availability of different sources of information brings the possibility to involve semantic evidence during the analysis of visual content in image collections, which is especially useful when considering the semantic gap [1], i.e., the discrepancy between visual features and semantic interpretations. Therefore, the combination of these data sources together with visual characteristics of images has received increasing attention from the research community in multimedia processing. The main problem is to take advantage of different data modalities to enable computer systems with the ability to make appropriate decisions according to the high-level task, which is known in the literature as multimodal fusion [2].

In this paper, we consider the problem of building a multimodal image representation that combines two data modalities: visual patterns extracted from images and text terms extracted from attached text descriptions. The proposed strategy mines the relationships between these two modalities to construct a unified representation based on Latent Semantic Analysis (LSA) principles. We propose a solution based on non-negative matrix factorization (NMF) to construct a latent-factor-based representation that can be spanned using text terms or visual features. We formulate a set of NMF-based algorithms for multimodal image analysis, which generates a joint visual–textual representation that is useful to approach different image analysis tasks.

The main contribution of our work is an NMF-based model to index multimodal data. In this work, multimodal collections are composed of images and some associated text descriptions. These image collections can be built from many different web sources, including Flickr and Picassa, in which several text descriptions for images can be identified using information extraction techniques. More details about the representation and preprocessing of each separate data modality are given in Section 3. Then, given a database of images with the corresponding text annotations, the multimodal analysis is performed using NMF algorithms as depicted in Fig. 1.

\begin{thebibliography}
\bibitem{1} Reference 1
\bibitem{2} Reference 2
\end{thebibliography}
The NMF-based strategy generates the latent semantic space using both data modalities. The goal is to find a set of latent factors that explain the underlying structure of the collection and the relationships between multimodal features. The latent representation is computed using a training data set composed of objects exhibiting both modalities. New objects can later be projected to the latent semantic space even if they do not have both data modalities. In consequence, the multimodal latent semantic model can deal with images without text or text without images, which is particularly useful to address several image analysis and retrieval tasks. The proposed models and their properties are presented in Section 4.

We evaluate the proposed strategy on two different tasks to demonstrate the potential of the multimodal representation: image indexing and automatic image annotation. Our method projects the input visual features to the multimodal latent semantic space to allow the subsequent analysis. Thus, in addition, we develop a set of algorithms for image search and image auto-annotation that are presented in Section 5.

An experimental evaluation was conducted using two image collections: Corel 5k [3], a collection of photographs with several tags and categories, and MIRFlickr 25000 [4], a data set of images downloaded from Flickr.com with the corresponding user generated tags and some additional labels provided as ground truth. The experimental setup and results are presented in Section 6, which shows that our proposed model outperforms baseline strategies. The final discussion and concluding remarks are presented in Section 7. Portions of this work have been previously reported in [5,6].

2. Relation to previous work

The use of multiple data modalities for multimedia analysis has become an important research topic during the last years. A comprehensive survey of the many research aspects of multimodal fusion for automatic multimedia analysis can be found in [2], which includes applications in audio, image and video processing using multiple data sources to achieve semantic decisions. Also, in the particular field of image retrieval, Datta et al. [7] discussed the importance of multimodal fusion for image indexing. The construction of systems that make semantic decisions using heterogeneous data sources is the ultimate goal of multimodal fusion.

Two main strategies can be considered for combining multimodal information: late fusion and early fusion. Late fusion, also known as rank aggregation or fusion at a decision level, consists in processing each data source separately during the indexing phase, with the multimodal integration taking place during the query phase. The work of Ah-Pine et al. [8] is an example of similarity combination to achieve multimodal access in image collections, using pseudo-relevance feedback to re-rank images in different applications. On the other hand, early fusion, or fusion at a feature level, consists in modeling feature relationships to create a new multimodal representation, so that during the decision phase, the only task to do is usually analyzing multimodal features [9,10]. Our work is categorized as an early fusion strategy for multimodal image analysis.

Latent topic analysis has been used to model the relationships between multimodal data, specifically images and text annotations. A set of generative models that use latent variables have been proposed to predict missing captions given unlabeled images [11,12]. These works are based on extensions of the latent Dirichlet allocation (LDA) model, in which a set of hidden factors are assumed to explain the associations between the two data types. Later, Monay and Gatica-Perez [13] proposed a simplified aspect model based on probabilistic latent semantic analysis (PLSA) to index and annotate images by jointly processing visual features and text data.

More recent works follow a latent topic analysis using matrix factorization approaches. Hare et al. [14] proposed a linear algebraic technique based on singular value decomposition (SVD) to learn a semantic space for image features and textual descriptions. This method is a multimodal extension of latent semantic indexing (LSI) for image retrieval that results in a semantic space suitable for image search. Latent topic analysis using matrix factorization has recently drawn of wide interest in information retrieval and image analysis. In particular, NMF algorithms have been used to analyze visual data to discover object classes [15] and to find correlations between image tags [16]. Other applications of NMF decompositions for visual data include [17–19].

All past works are different from ours since they are focused on processing either visual features or text annotations rather than exploiting multimodal interactions between both data types. Our work is the first one, according to our knowledge, that addresses the multimodal indexing problem using an NMF-based algorithm. In [5,6], we addressed the problems of multimodal image indexing and automatic image annotation, respectively. The present paper builds upon these works by proposing a unified method for solving both problems, and performing a systematic and extended experimental evaluation.

3. Multimodal image collections

Assume an image collection with attached unstructured text annotations. An excerpt of text may be identified from the source document for each image using information extraction techniques to locate captions, to parse image names and tool-tips, among others [20]. After the information extraction step, each image has an associated unstructured set of text terms. In our model, it is
not required that every image has a text description, since it is reasonable to find many of them without surrounding text, or it might be difficult to identify a reliable piece of text to be attached. However, it is assumed that there are enough annotated images to perform a multimodal analysis. These collections containing portions of images without text descriptions will be referred to as partially annotated image collections, and that portion of the database will be indexed later after the multimodal analysis has taken place (Fig. 2).

We assume the availability of a large enough sample of images taken from the collection together with their associated text descriptions extracted from the source documents. This sample will be called the training data set. Hence, the first step of multimodal analysis is to prepare each individual data modality separately in the training set. This preprocessing step aims to build two matrices, one with visual data and the other one with text data. Each image is represented as a vector in the visual matrix, while its corresponding text description is represented as a vector in the text matrix. In principle, any vector representation for images and text descriptions is allowed as long as all their features are non-negative. Then, the matrices $X_v$ for visual features and $X_t$ for text features will be non-negative matrices as well.

Different representations for visual image contents are naturally non-negative [21]. In this work, we adopt a bag-of-features approach to represent image content. We extract blocks of $8 \times 8$ pixels from a set of training images with an overlap of 4 pixels along the x- and y-axes to build a set of training blocks. Each block is processed in the three RGB color channels using the discrete cosine transform (DCT) and the 21 largest coefficients per channel are used as features, leading to a block descriptor of 63 features. Then, the matrix $X_v$ is processed in the three RGB color channels using the discrete cosine transform (DCT) and the 21 largest coefficients per channel is processed in the three RGB color channels using the discrete cosine transform (DCT) and the 21 largest coefficients per channel.

A bag of features $X_v$ is constructed for images and text terms is allowed as long as all their features are non-negative. Then, the matrices $X_v$ for visual features and $X_t$ for text features will be non-negative matrices as well.

The purpose of a latent factor model is to try to explain these occurrences by characterizing both images and features using a set of $f$ factors inferred from occurrence patterns. For images, discovered factors might measure dimensions such as natural scenes versus buildings, or amount of forest cover as opposed to man-made objects; less well-defined dimensions such as illumination conditions or distance to focused objects; or completely uninterpretable dimensions. For features, each factor measures how much a feature appears in images related to the corresponding factor. We assume these factors to form new meaningful dimensions to organize images in the collection. Fig. 3 illustrates this idea for a simplified example with two dimensions.

4. Multimodal latent factor analysis

Each of the two matrices described above, $X_v$ and $X_t$, provide information about the occurrence of different features for each image in the collection. The concept of occurrence is borrowed from the bag-of-words and bag-of-features models, in which values account for the number of times a particular element appears within the image. In general, these values may be understood as occurrence ratios between features and images, so each image is characterized by the occurrence of certain features and each feature is characterized by the images in the collection in which they appear.

In our model, images, visual features and text terms can be represented together in a joint latent factor space, their relationships can be explained using the inner product between their corresponding representations. Let $h_i \in \mathbb{R}^f$ be the representation of an image $i$, and $w_f \in \mathbb{R}^f$ be the representation of a feature $f$, both in the latent factor space. The elements of $h_i$ measure the extent to which the image $i$ expresses latent factors. The elements of $w_f$ measure the extent to which the feature $f$ appears in images.
associated to the corresponding factors. Thus, the resulting dot product between these representations models the occurrence of the feature \( f \) within the image \( i \), as follows:

\[
x_{ij} = w_h^j h_i = \sum_{k=1}^r (w_k^j)(h_k)_i
\]  

From an image collection point of view, the occurrence patterns can be expressed using matrix notation in the following way:

\[
X = WH 
\]  

where \( X \in \mathbb{R}^{p \times l} \), \( W \in \mathbb{R}^{p \times r} \), \( H \in \mathbb{R}^{r \times l} \), \( p \) is the total number of available features, \( r \) is the number of latent factors and \( l \) is the number of images in the collection. Notice that both features and images have a vector representation using \( r \) latent factors in the matrices \( W \) and \( H \) respectively. Also, the matrix \( W \) is considered as the basis of the latent space, since each image in \( X \) is represented through a linear combination of \( W \)’s columns using the coefficients in \( H \). Thus, the main problem is to compute these representations out of the original feature matrix \( X \), i.e., to find a factorization of \( X \) in terms of \( W \) and \( H \).

### 4.1. Latent factors via singular value decomposition

A common approach to compute the latent factors in information retrieval is using singular value decomposition (SVD). This strategy consists of estimating a rank-reduced factorization of the feature matrix in terms of its eigenvectors and eigenvalues,

\[
X = U\Sigma V^T 
\]  

where \( U \) and \( V \) are orthonormal matrices. To generate a semantic space using this decomposition, the eigenvalues in \( \Sigma \) are sorted in decreasing order to preserve the first \( r \) largest eigenvalues while the rest are set to zero. This is equivalent to dividing the matrices of the decomposition as \( U = [U_r U_c] \); \( V^T = [V_r V_c]^T \) and splitting \( \Sigma \) into two matrices, \( \Sigma_r \), a squared matrix with the first \( r \) selected eigenvalues and \( \Sigma_c \) with the remaining eigenvalues. Then, the matrix factorization can be rewritten as

\[
X = U\Sigma_r V_r^T + U_c\Sigma_c V_c^T 
\]

Assuming that \( X \) has \( r \) independent factors, it can be shown that the best rank-\( r \) approximation to \( X \), in the least squares sense, is given by \( X_r = U_r\Sigma_r V_r^T \). Using this low rank approximation to \( X \), the basis of the latent factor space is \( W = U_r \) and the latent image representation is \( H = \Sigma_r V_r^T \).

Under this scheme, the basis of the latent factor space is composed of a set of orthogonal vectors from the matrix \( U \). The criterion used to select the set of vectors is based on the size of the corresponding eigenvalues in \( \Sigma \), since the larger the eigenvalue, the larger the feature variance of the collection in that direction, and the better the low-rank approximation. In that sense, the latent factors obtained using SVD are orthogonal factors that maximize the variance of the data representation.

### 4.2. Latent factors via non-negative matrix factorization

The observed occurrence values in the feature matrix \( X \) may be modeled directly by learning the factor matrices \( W \) and \( H \) for features and images respectively, using alternative matrix factorization techniques to SVD. Two main requirements are herein considered to approximate the matrix factorization in Eq. (2).

First, the resulting basis for the latent factor space should allow non-orthogonal vectors as well as orthogonal ones, as long as they correspond to structural patterns in the feature matrix. Second, the matrix approximation must allow non-negative values only, for both the basis vectors and the codifying vectors.

Notice that an approximation of the matrix factorization is allowed rather than requiring an exact matrix factorization. Thus the matrix factorization in this problem, may be expressed as \( X \approx WH \), or more precisely, \( X = WH + E \), where \( E \) is a matrix of approximation errors. Solving this approximation can be approached as an optimization problem to find \( W \) and \( H \) that minimizes the Frobenius norm of the error matrix \( ||E||^2 = ||X-WH||^2 \), that is, the difference between the original matrix and its approximation.

Another objective function to evaluate the factorization approximation is the Kullback–Leibler (KL) divergence:

\[
D(X||WH) = \sum_{ij} X_{ij} \log \frac{X_{ij}}{(WH)_{ij}} X_{ij} + (WH)_{ij}. 
\]

This objective corresponds to the KL divergence between the empirical distribution of features in the matrix \( X \) and the model distribution \( WH \). This amounts to projecting the observed occurrences on the subspace spanned by the factors based on the KL-divergence [23]. This is different from a squared error based on Frobenius norm which would result in an orthogonal projection, resulting in a more appropriate model to deal with the data representations considered in this work: bag-of-features histograms for both visual and textual content, which can be interpreted as probability distributions.

According to Liu et al. [24], minimizing the KL divergence is obtained by maximizing the likelihood of observing the data matrix under the assumption that it follows a Poisson Model. Minimizing the Frobenius norm is obtained by maximizing the likelihood of observing a Gaussian distributed data matrix. The Poisson distribution offers a more faithful model, compared to the Gaussian distribution, for representing “counts”, which is exactly the case for our bag-of-words and bag-of-features representation for the text and visual content of our multimodal data, respectively.

Therefore, the formulation of the optimization problem to approximate a non-negative matrix factorization (NMF), using the divergence objective function, is formally denoted as

\[
\min_{W,H} D(X||WH) \quad \text{s.t.} \quad W,H > 0 
\]
This optimization problem is convex in \( W \) only or \( H \) only, but it is not convex in both variables together. However, there are different techniques from numerical optimization that can be applied to obtain a good approximate solution. In particular, an algorithm to find \( W \) and \( H \) simultaneously based on multiplicative updates, has been shown to yield non-increasing objective values and to converge to stable solutions [25]. These multiplicative update rules are as follows:

\[
H_{\text{old}} \leftarrow H_{\text{old}} \frac{\sum \alpha \nabla H_{\text{old}} X_{\text{old}} (WH)_{\text{old}}}{\sum \alpha \nabla W_{\text{old}} H_{\text{old}}}
\]

(7)

\[
W_{\text{old}} \leftarrow W_{\text{old}} \frac{\sum \alpha \nabla W_{\text{old}} X_{\text{old}} (WH)_{\text{old}}}{\sum \alpha \nabla W_{\text{old}} H_{\text{old}}}
\]

(8)

Computing \( W \) and \( H \) in this way provides a good tradeoff between ease of implementation and speed of convergence. It is straightforward to see that the multiplying factors used to update matrices \( W \) and \( H \) become equal to 1 when \( X = WH \), which means that perfect factorization is necessarily a fixed point of the update rules. The proof of convergence for this strategy was provided by Lee and Seung [25] using an auxiliary function that is similar to the one used to prove the convergence for the Expectation Maximization algorithm.

The factorization is set to decompose the input matrix into a fixed number of factors defined by the parameter \( r \). In our setup, the number of factors (that defines the complexity or order of the factorization model) is found using cross validation on the training data to compare the performance of models with different complexity. Tan and Fevotte [26] presented a framework to construct multimodal image representations are proposed as a way to generate semantics space using visual features instead of text terms, as the former is a matrix whose rows are indexed by \( n \) visual features and whose columns correspond to the \( l \) images in the database. The latter has \( m \) rows to represent text terms and \( l \) columns for images as well. The construction of a latent semantic space may be done by decomposing the matrix of images that have only visual features or only textual annotations. However, to generate a semantic space for image indexing, we are interested in exploiting multimodal relationships. Thus, two strategies to construct multimodal image representations are proposed as follows.

4.3. Multimodal latent factors

The image database is composed of two data modalities, herein denoted by \( X_0 \in \mathbb{R}^{n \times l} \) and \( X_1 \in \mathbb{R}^{m \times l} \). The former is a matrix whose rows are indexed by \( n \) visual features and whose columns correspond to the \( l \) images in the database. The latter has \( m \) rows to represent text terms and \( l \) columns for images as well. The construction of a latent semantic space may be done by decomposing the matrix of images that can have only visual features or only textual annotations. However, to generate a semantic space for image indexing, we are interested in exploiting multimodal relationships. Thus, two strategies to construct multimodal image representations are proposed as follows.

4.3.1. Mixed multimodal representation

This strategy consists of the construction of a multimodal matrix \( X = [x_0 X_1 \quad (1 - x) X_1^T] \), with \( x \in [0, 1] \), a weighting parameter that controls the relative importance of the two data modalities. We set \( x = 0.5 \) in our experiments (to give the same importance to visual and text data), unless otherwise stated. Then, the matrix is decomposed using NMF as follows:

\[
X_{\text{old}} = W_{\text{old}} H_{\text{old}}
\]

(9)

where the subindices indicate the dimensions of the matrices, \( W \) is the basis of the latent space, i.e., the latent factors, in which each multimodal object is represented by a linear combination of the \( r \) columns of \( W \). These factor proportions are codified in the columns of \( H \). The mixed multimodal representation aims to find correlations between features of both modalities, i.e., to find relationships between visual features and text terms, since both of them are aligned in the same feature matrix. A similar approach using SVD was proposed by [27], in which visual features and text terms are aligned to generate a multimodal latent semantic representation.

4.3.2. Asymmetric multimodal representation

The previous strategy decomposes the multimodal information by building a multimodal matrix with visual features and text terms. It has been reported in the literature that text descriptions tend to provide a more reliable information source to extract semantic information for image retrieval than visual features [13]. Evidence of this fact can be observed in different image retrieval challenges that provided data sets with images and text descriptions, and final polls show a dominant position of text-based approaches [28,29]. Thus, we present an asymmetric approach for the construction of the latent semantic space that first derives a semantic image representation from text data, and then follows an adaptation of the visual representation to fit the semantic one. In other words, the text information plays the role of a leader compared to the visual content. However, both modalities are exploited to discover the semantic representation.

The algorithm has two main steps to construct the semantic space:

1. **Building a semantic image representation**: This step decomposes the text matrix using the NMF algorithm:

\[
X_t = W_t H_t
\]

(10)

In this case, the \( m \times r \) matrix \( W_t \) contains the vectors of a basis in which text terms are correlated with \( r \) latent semantic factors. After this step, the semantic representation for all images in the data set is codified in the matrix \( H_t \).

2. **Adapting the latent space basis**: To complete the basis of the multimodal latent space, the construction of a basis for visual features is adapted to match the previously obtained semantic representation. That is, we find a matrix \( W_{v, n \times l} \) which spans the semantic space using visual features instead of text terms, as follows:

\[
X_v = W_v H_v
\]

(11)

Notice that in the second step, the matrices \( X_t \) and \( H_t \) are already known, then, the problem of finding \( W_v \) may be accomplished by computing the multiplicative updates for \( W \) while fixing \( H \). We refer to this computation step as the adaptation algorithm in the rest of the paper.

The convergence of this modified algorithm to obtain the factorization \( X_v = W_v H_v \) can be understood by analyzing the optimization problems in each step. In the first step of this algorithm, the latent representation for training images is obtained in \( H_t \) running update rules in Eqs. (7) and (8), which have been shown to converge to a local minimum [30,25]. In the second step of our asymmetric approach, the matrix \( H_t \) is fixed, and we solve it by running the update rule in Eq. (8) only. It is known that by fixing one of the matrices in the factorization makes the problem of finding the unknown convex [25], thus a global minimum can be found. However, this will be a global minimum with respect to the previously found matrix, which is a local minimum of the first problem.

5. Image indexing and auto-annotation

5.1. Image indexing

The main goal of image indexing is to generate an image representation that can be used to match similar contents given a
particular information need. After the multimodal decomposition has been done using the algorithms described above on a training set, all other images in the collection have to be indexed, i.e., all of them have to be projected to the latent semantic space. Consider a partially annotated image collection, in which images with and without associated text can be found. Also, since the system is based on a multimodal index, we can consider three different ways to query the system: using example images, using keywords, and using both. Be it for indexing images or to process queries, we can project all data to the latent space using the following strategies:

<table>
<thead>
<tr>
<th>Input data</th>
<th>Projection</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y ): multimodal</td>
<td>( y = Wh )</td>
</tr>
<tr>
<td>( y_v ): visual</td>
<td>( y_v = W_vh )</td>
</tr>
<tr>
<td>( y_t ): text</td>
<td>( y_t = W_th )</td>
</tr>
</tbody>
</table>

Any input data has to be projected to the latent space by finding \( h > 0 \) in terms of the corresponding basis of the multimodal space. When multimodal data is available, the full factor basis matrix \( W \) is used. When only visual or text data is available, the corresponding sub-matrix of the basis (\( W_v \) for visual data, or \( W_t \) for text data) is used. The projection of \( y \) on the latent space, \( h \), is found by using the multiplicative updating rules for \( h \) while keeping the matrix \( W \) fixed. We refer to this computation step as the codification algorithm, since new data and the basis of the latent space are given to find the multimodal latent representation.

Once images and queries have been projected onto the multimodal latent space, a similarity measure is needed to identify relevant results. We use the dot product as similarity measure, and results are ranked in decreasing order of similarity. Notice that the dot product in the multimodal latent space gives a notion of the extent to which two vectors share similar values for latent factors.

### 5.2. Image auto-annotation

The problem of image auto-annotation is to assign a set of keywords to an image that do not have any attached text. As in the case of image search using visual queries, we need to project images to the multimodal latent space using the codification algorithm described in the previous subsection. Then, to find the annotation words that best describe an unannotated image, we first need to compute its similarity to training images as follows:

\[
 z = h^T H
\]

(12)

where \( h \) is the latent representation of the unannotated image, \( H \) is the representation of the training images in the semantic space, and \( z \in \mathbb{R}^k \) is a vector of similarity scores. To output annotation words to the query image, we compute their counts in the set of most similar training images. Then the annotation words are ranked according to their counts. This step is similar to what has been proposed in other works \([31,32]\) and is useful here to assess the benefits of using the NMF latent space construction compared to other multimodal-based methods.

### 6. Experimental evaluation

#### 6.1. Datasets

##### 6.1.1. Corel 5k data set

The Corel 5k image database is composed of 5000 images in 50 categories and has been manually annotated using a text vocabulary of 371 terms \([3]\). This data set has been used as a benchmark in automatic image annotation and retrieval research, allowing the comparison of several strategies during the last years \([14,33,13,34]\). The bag-of-features approach is used to represent visual contents using a dictionary of 2000 visual patterns. The conventional experimental protocol is followed in this work, regarding testing and training partitions. The data set was split into three subsets with 4000 images for training, 500 images for validation and 500 for final tests. The training set is used to feed the learning algorithms while the validation set is used for parameter tuning.

#### 6.1.2. MIRFlickr 25000 data set

The MIRFlickr-25000 image data set is composed of 25,000 pictures downloaded from the popular online photo-sharing service Flickr. These photos were collected directly from the web, to provide a realistic data set for image retrieval research, with high-resolution images and associated metadata \([4]\). The data set comes with the Flickr tags given by users, which can be considered as low level, noisy text. By processing this content, a 2105-word dictionary is defined based on the most frequent terms. The bag-of-features approach is used to represent visual content using a dictionary of 1000 visual patterns. This image collection has also been manually annotated using a set of 38 semantic terms provided as ground truth. The annotation vector has binary elements indicating whether the photo can be described by the term or not. These are considered as high-level textual descriptions.

We follow the conventional training–validation scheme, using 15,000 images for training and the remaining 10,000 images for testing \([35]\). For the retrieval experiments, 1000 random images are taken as queries from the test set. Thus, the database for these experiments contains 24,000 images, from which 15,000 have text annotations and the remaining 9000 have only visual features. This is a quite realistic setup, in which the image collection is partially annotated, and the multimodal analysis needs to be generalized to the remaining images.

#### 6.2. Visual indexing

We evaluate the response of an image retrieval system that uses only visual features to retrieve images from the database when a visual query is provided. Our content-based descriptor is based on the bag-of-features approach, so images are represented by histograms of the occurrence of visual patterns in a codebook. To match image content using this representation, the histogram intersection is used as a similarity measure.

In addition to this visual matching strategy, we evaluate the performance of the system using a latent semantic space built upon visual information only. The NMF and SVD algorithms are fed with the matrix of visual descriptors and the search is performed in the semantic space using the dot product as similarity measure (essentially the same as cosine since the data is normalized to unit 1). To evaluate the performance of the retrieval response, the mean average precision (MAP) is computed in each experiment. In the Corel 5k data set, an image in the results is considered relevant if it shares the same category with the query. For the MIRFlickr data set, an image in the results is considered relevant if it shares at least one semantic label with the query.

Experimental results on the Corel data set showed that direct matching performs better than using a visual latent space either with SVD or NMF \([5]\). The maximum performance using only visual data reaches a MAP value of 0.101. On the MIRFlickr data set, visual latent spaces perform slightly better than direct
matching, reaching a MAP value of 0.557. This strategy does not take advantage of the text annotations in the collection, and works as a baseline to measure the improvement of the proposed multimodal approach.

6.3. Multimodal indexing

The multimodal analysis is performed using the training data by applying three algorithms: SVD mixed, NMF mixed and NMF asymmetric. Afterwards, all images and queries are indexed in the latent factors’ space and the evaluation is carried out by observing the performance in terms of MAP. The experiments follow the Query by Example Paradigm (QBE), to evaluate the response of an image retrieval system that indexed images using multimodal data, even though the expected queries have only visual information. This evaluation challenges the algorithm's ability to retrieve semantically valid results in the absence of text annotations in the query.

Fig. 4 presents the performance on the validation sets for all indexing strategies for both data sets, using different sizes of the latent space. All multimodal strategies show the construction of improved indexes for image search based on the QBE paradigm. Overall, NMF-mixed and SVD multimodal presented very similar performance in both data sets. The results show that the proposed NMF-asymmetric indexing algorithm achieves a better performance with respect to all other models. This shows that the proposed strategies effectively involve text semantics in the organization of visual patterns using multimodal factors. Notice that in most of the cases, the number of latent factors required to observe an improved performance is relatively low (less than 100 for both collections).

Table 1 shows MAP values for all the evaluated models, computed on the test sets in both collections. Experiments on the Corel data set show large improvements when using a multimodal index to search with images without text data. The relative improvement in the MIRFlickr data set is modest compared to the one observed in the Corel data set. This is because the MIRFlickr is an image collection extracted from a real online service, hence it is more noisy and challenging. This provides more realistic conditions and is also a bigger data set with a ground truth given by multiple semantic labels rather than clearly defined categories. Despite all these differences, the proposed strategies show better performance with respect to baseline models, resulting in an improved retrieval response.

6.4. Weighted multimodal indexing

In Section 4.3.1 the NMF-mixed algorithm was introduced with a weighting parameter that controls the relative importance of visual and text data modalities. In this section we investigate the impact of alternative weightings to find the multimodal latent factors. The weighting parameter $\alpha$ allows a flexible configuration of the multimodal decomposition. When $\alpha = 1.0$, only visual information is considered in the matrix factorization algorithm. When $\alpha = 0.0$, the text data is the only one used to build the latent factor space, which is basically the same as the NMF-asymmetric algorithm. Intermediate values for $\alpha$ may result in different performances according to the contribution of each modality.

Fig. 5 presents the performance response of the NMF-mixed algorithm when the weighting parameter $\alpha$ is changed. Both data sets, Corel 5k and MIRFlickr, show basically the same tendency: as long as more weight is given to the text modality, a better response in terms of MAP is observed. This is mainly due to the evaluation protocol followed for both data sets, which is based on a ground truth that relies on semantic categories or semantic labels. Text annotations are usually more correlated to these semantic representations, and might be considered closer to the human interpretations than visual features alone. However, we believe that under other evaluation scenarios, which may include more perceptual or subjective criteria, giving more weight to the visual modality may be of benefit to the underlying task. Examples of these alternative scenarios include exploratory image search [36] and visual pattern mining [37,38].

6.5. Answering multimodal queries

The previous sections have evaluated the response of the multimodal indexing system under the Query by Example paradigm, i.e.,
assuming that users express their information need using example images. That evaluation allows to assess the influence of multimodal data in the visual retrieval task, and enables the system to give more meaningful results when there is no other clue other than a visual example. This search mode may support the operation of online image search services for users with camera phones and other mobile devices that are used to capture an image and then look for similar ones.

However, the proposed multimodal retrieval system can handle other types of query paradigms, as was mentioned in Section 5.1, including query by keywords and even multimodal queries, i.e., queries with visual examples and text descriptions. Hence, the following experiments aim to evaluate other types of query paradigms to search in the multimodal index. We consider two main scenarios: first, an image collection that is fully annotated, in other words, text descriptions are available for every image in the database; second, a non-annotated image database, for which the multimodal analysis has been extended from a training sample. We used the NMF-asymmetric algorithm on these experiments on two subsets of the Corel 5k data set (fully annotated and non-annotated).

Table 2 reports MAP figures for this evaluation. Visual-only queries refer to the QBE paradigm evaluated in the previous sections. Notice that when these visual queries are enhanced with some keywords to provide a multimodal query, the system response is improved. This is consistent with the notion that, as long as users provide more clues about their information need, the system should be able to retrieve more relevant results. In the case of text-only queries the situation is different regarding the type of collection that has been queried. On a fully annotated collection, the multimodal index performs better using keywords-only than using a multimodal query. However, when the collection do not have any text annotation, multimodal queries work slightly better than keywords-only.

This shows that the multimodal index is able to support multiple query paradigms even if the image collection is partially annotated. Then, other methods that rely on visual content only, would provide poor information because of the semantic gap, and methods that rely on text content only, would leave large portions of the images inaccessible. The proposed multimodal index can handle all these situations in a unified fashion.

6.6. Image auto-annotation

Automatic annotation is performed in the proposed framework by searching similar images in the latent factors’ space and selecting frequent terms associated to the top results. Auto-annotation experiments were carried out using the training-validation scheme to tune up parameters (number of factors and number of nearest neighbors). Annotations on the Corel 5k data set are chosen from the 371 terms of the attached text [3,14,33,13]. Annotations on the MIRFlickr data set are chosen from all 38 semantic labels (relevant and potential) [35,4]. Performance is measured using standard precision scores that indicate how many relevant tags have been assigned to query images.

Table 3 reports annotation performance in terms of MAP for both data sets, using the three multimodal strategies. These results are from the best performing configurations and show that NMF-asymmetric has a better performance than the mixed strategies. The annotation performance has a direct relationship with the retrieval performance since the underlying approach to assign terms is based on searching similar images. However, the appropriate factorization changes significantly according to the set of annotation terms: notice that the number of factors required to achieve the best performance is related to the number of annotation terms in the case of the asymmetric strategy.

Table 2

<table>
<thead>
<tr>
<th>Database</th>
<th>Query type</th>
<th>Visual</th>
<th>Multimodal</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully annotated</td>
<td></td>
<td>0.2289</td>
<td>0.3345</td>
<td>0.3746</td>
</tr>
<tr>
<td>Non-annotated</td>
<td></td>
<td>0.1709</td>
<td>0.2211</td>
<td>0.2205</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Model</th>
<th>Corel 5k</th>
<th>MIRFlickr</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVD mixed</td>
<td>0.2663</td>
<td>0.5617</td>
</tr>
<tr>
<td>NMF mixed</td>
<td>0.2948</td>
<td>0.5775</td>
</tr>
<tr>
<td>NMF asymmetric</td>
<td>0.3180</td>
<td>0.6215</td>
</tr>
</tbody>
</table>

Fig. 5. Effect of the weighting parameter, $\alpha$, on the performance of the NMF-mixed algorithm. The weighting parameter controls the contribution of the visual and text data modalities to the initial construction of the latent space. $\alpha=0$ corresponds to the asymmetric multimodal representation, which is built exclusively using text data, and greater values indicate a greater contribution of visual modality. (a) Corel 5k data set. (b) MIRFlickr data set.
The complexity increases with the number of features, and the time complexity of the algorithm is 

\[ O(p \times l \times r \times k) \]

where \( p \) is the number of training examples, \( l \) is the number of latent factors. Notice that the complexity increases with the number of features \( p \), the number of training examples \( l \) and the number of latent factors \( r \).

In practice, matrix multiplication can gain substantial performance using parallelization due to the nature of these operations. Our implementation was developed in Matlab and run on a server machine with 32 GB of RAM memory and eight CPU cores. The largest matrix factorization in our experiments was done for the Corel 5k data set, using 15,000 training examples, 5000 features, and the number of latent factors \( r \). In average, this decomposition takes 15.2 min using all cores and 3 GB of memory. In the case of the Corel 5k data set, the largest decomposition involving 4500 training examples, 2300 features, and 500 latent factors, takes about 3.7 min in average.

In search time, every image has already been indexed in the latent factor space, which in our experiments, turned out to be small, between 40 and 100 factors. A similarity search using the dot product is linear in the number of images on the database, which takes several milliseconds using our implementation in Python. This search can be also scaled up easily to compute similarities in parallel and using advanced indexing techniques.

7. Discussion

7.1. Multimodal representations

Joint image-text analysis has been of wide interest since the seminal work of Barnard et al. [11]. Most of the research efforts have been oriented to enable text-based image retrieval or to predict relevant annotations [3,33,39,27,13,40,14,10,35]. Our work is oriented to build a multimodal representation for images that integrates visual features and text terms for solving a variety of tasks. In most of our experiments, we used the text modality as an auxiliary source of information rather than using it as the target element to be predicted or explained by the models. The introduction of text information in the image representation was shown to provide important improvements in all the experimental setups presented in this work, demonstrating the potential of the proposed approach.

Our model provides a unified framework to deal with multimodal representations useful to approach a wide variety of tasks for image indexing and search. Few works have explored other tasks beyond keyword-based search, such as querying using semantic examples and pictures from mobile phones [41,42,34]. We do not restrict our evaluation to keyword-based search or semantic examples, instead we demonstrated the potential of multimodal representations for image indexing on fully and partially annotated collections, for searching with different query paradigms and for performing auto-annotation. We believe that a truly multimodal image management system should support all these diverse capabilities in an integrated framework and that is precisely one of the contributions of our work.

The evaluation carried out in this work was mainly oriented to measure the performance of the multimodal representation to make semantic decisions, i.e., retrieve relevant results or provide high-level annotations. However, the proposed multimodal representation may be of great benefit for other tasks such as perceptual image analysis or subjective evaluations, including visual pattern mining, image collection visualization and exploratory image search.

7.2. Latent factors via NMF

We investigated the potential of matrix factorization to build multimodal latent factors to represent images. Latent factors can be obtained using a variety of methods including latent semantic indexing (LSI), probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA). We discuss some differences and similarities between NMF and these other three approaches, that have also been used to model joint image-text data sets.

Hare et al. [27] used an LSI-like approach to build semantic spaces using visual and text data. We consider that modeling orthogonal factors is a restrictive strategy, since some latent factors might be related to each other, indicating meaningful patterns in the collection [43]. In addition, the basis and the latent representation are not restricted in terms of the sign that their dimensions can have, that is, latent factors may have positive or negative values.
negative components. Different image and text representations are naturally non-negative as was discussed in Section 3. Thus, from a data analysis viewpoint, it is reasonable to model the structure of the collection using non-negative representations as well. Lee and Seung [30] demonstrated how to learn parts of objects by restricting the element representations to be non-negative. These properties lead to a more meaningful image representation, since latent factors can be seen as meaningful parts that can be combined in an additive way to understand image content.

Monay and Gatica-Perez [13] proposed modeling semantic aspects using PLSA to integrate both visual and text data. Although NMF and PLSA have been shown to optimize the same objective function, Ding et al. [44] emphasize the fact that they are different algorithms and converge to different solutions. PLSA deals with data from a statistical viewpoint, using maximum likelihood estimation to find an approximate latent representation. NMF models data from a sub-space viewpoint, using optimization strategies to approximate the matrix decomposition. Therefore, multimodal factors in PLSA consist of probabilities associated to features and terms, whereas NMF provide multi-modal factors as vectors with visual features and terms which again, can be seen as meaningful parts of objects in the image collection.

Blei and Jordan [12] modeled the joint distribution of text terms and image segments using LDA, based on a generative model in which the visual data is the primary modality and is generated first. Thus, conditioned on the topics used for an image, text terms are then generated. This design explains the observed data following the process of an annotator, in which images are observed and then annotated. In most of our experiments, the NMF-asymmetric strategy performed better than NMF-mixed, suggesting that the text data should be used as the primary modality to approach semantic decisions (see previous section for a discussion about evaluation). A generative model for our NMF-asymmetric algorithm would generate text data first and then the visual one. This design would explain the data following a painter process, in which the text is given and the pictures are then painted. This difference between Blei and Jordan’s model and ours, motivates further research to better understand joint image-text modeling.

8. Conclusions

We presented an approach for building multimodal image representations using non-negative matrix factorization as a method to create latent semantic factors where data of different modalities can be associated. The two data modalities involved in our work are visual features and text terms. The main goal of the proposed multimodal representation is to reduce the semantic ambiguity of the visual content. Since we have experimentally found that textual data is more reliable for building topics of images, we introduced the asymmetric NMF method that exploits text data first and then adapts the basis of the visual descriptors. This can be seen as an enforcement of visual patterns to be organized according to the text semantics.

The experimental evaluation showed the potential of the proposed multimodal representation to approach a wide variety of tasks associated to image indexing in a unified framework. We demonstrated how to build multimodal factors, to extend them to partially annotated collections, to search using multiple query paradigms and to annotate images automatically, all supported by the same core methodology. All the evaluations were carried out using two standard real data sets, and the proposed approach consistently performed better than baseline methods. This work shows that matrix factorization strategies can be effectively used to model multimodal latent factors and also that multimodal representations are very useful to perform multiple image collection analysis tasks.

Acknowledgments

This work was partially funded by the project REPRESENTACIÓN Y CLASIFICACIÓN DE GRANDES COLECCIONES DE IMÁGENES MÉDICAS number 1101-489-25777 of Colciencias by Convocatoria COLCIENCIAS 489 de 2009. This work was also partially funded by the US National Science Foundation Grant number 0916489, and by the project SISTEMA DISTRIBUIDO DE ANOTACIÓN AUTOMÁTICA Y RECUPERACIÓN SEMÁNTICA DE IMÁGENES DE HISTOLOGÍA number 1101-487-25779 of Ministerio de Educación Nacional de Colombia by Convocatoria COLCIENCIAS 487 de 2009.

References


[34] Jaafar Ben-Abdallah is a Ph.D. candidate in Computer Science at the Department of Computer Systems and Industrial Engineering, National University of Colombia. His research focuses on machine learning and image retrieval techniques applied to multimodal data collections. In 2010 he was a research intern at Microsoft Research, working at the Interactive Visual Media group. He is currently a research intern at Google, working at the Image Search group.

[35] Fabio A. González is an Associate Professor at the Department of Computer Systems and Industrial Engineering, National University of Colombia. He earned M.Sc. and Ph.D. in Computer Science degrees from the University of Memphis in 2003 and an M.Sc. in Mathematics from the National University of Colombia in 1998. His research areas include machine learning, information retrieval and computer vision. He has more than 60 refereed publications including journal papers, an edited book, book chapters and conference papers. In 2009, Dr. González received the prestigious Fulbright Visiting Scholar Fellowship.

[36] Olfa Nasraoui is the Endowed Chair of e-commerce and the founding director of the Knowledge Discovery and Web Mining Lab at the University of Louisville. She received her Ph.D. in Computer Engineering and Computer Science from the University of Missouri-Columbia in 1999. From 2000 to 2004, she was an Assistant Professor at the University of Memphis. Her research activities include Data Mining, Web mining & Personalization, and Computational Intelligence. She is the recipient of the National Science Foundation CAREER Award for outstanding young scientists, and the recipient of the best paper award in theoretical developments in computational intelligence at the Artificial Neural Networks in Engineering conference (ANNIE 2001). She has more than 110 refereed publications, including 35 journal papers and book chapters and eight edited volumes. Her research is funded mainly by the US NSF and by NASA. She has served as a reviewer, program Vice Chair and organizer for several data and Web mining conferences and workshops. Between 2004 and 2008, she has organized the WebKDD workshops on Web Mining and Web Usage Analysis, held each year in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).