Online Matrix Factorization for Multimodal Image Retrieval

Juan C. Caicedo and Fabio A. González

Universidad Nacional de Colombia

Abstract. In this paper, we propose a method to build an index for image search using multimodal information, that is, using visual features and text data simultaneously. The method combines both data sources and generates one multimodal representation using latent factor analysis and matrix factorization. One remarkable characteristic of this multimodal representation is that it connects textual and visual content allowing to solve queries with only visual content by implicitly completing the missing textual content. Another important characteristic of the method is that the multimodal representation is learned online using an efficient stochastic gradient descent formulation. Experiments were conducted in a dataset of 5,000 images to evaluate the convergence speed and search performance. Experimental results show that the proposed algorithm requires only one pass through the data set to achieve high quality retrieval performance.

1 Introduction

Consider the problem of finding useful images by querying a system with an example image, i.e., the user provides an image to retrieve other semantically related images from a large collection [1]. This kind of search —also known as the query-by-example paradigm for image retrieval [2]— can be of potential benefit in different situations such as people taking pictures with mobile phones [3], or physicians comparing a new medical image with respect to the hospital's archive [4]. Regardless of the specific task, the main problem of finding useful images with an example query is the semantic gap [2]: images with similar visual features computed by a machine can have different semantic meanings for users observing them.

Then, the main research agenda in image retrieval systems has been devoted to design methods able to automatically understand image contents, and so, image search systems will have the ability to deliver more accurate results. The most popular approach is based on auto-annotation, which consists on analyzing the visual content of an image and generating tags associated to what can be seen on it [5]. However, one of the main drawbacks of this approach is scalability: most of these methods only work with a few tens of labels or tags. For large scale image search systems, the ability to handle large numbers of tags would be a desirable property. Also, methods for large scale image search are required to handle large amounts of images, which can be very expensive in computational terms.

In this paper, we explore the ability of matrix factorization algorithms to build large scale multimodal image indices. First, the image collection is organized in a table of images vs. features, which encodes the visual contents of images. This table is actually a

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 340–347, 2012.

matrix of visual data, that can be analyzed and processed to extract meaningful patterns. Second, a similar matrix is build for text data associated to training images, i.e., a matrix of images vs. tags, which can handle any number of keywords simultaneously.

The goal of the method proposed in this paper is to find relationships between these two matrices, using multimodal analysis. We propose the use of matrix factorization algorithms to decompose a training data set, and find correspondences between visual patterns and text terms. These correspondences are expressed as latent factors that can be computed from the input matrices described before. Under this formulation, a system for image search can directly handle multiple tags or labels for a single image, and can handle new images that do not have annotations at all. Furthermore, the proposed decomposition algorithms are formulated using stochastic gradient descent, which allows to manipulate very large image collections.

The main contribution of this paper is the formulation of the matrix factorization algorithms as online processes and its application to multimodal image representation. The contents of this paper are organized as follows: Section 2 reviews related work. Section 3 discusses the visual representation and similarity measures for image search. Section 4 presents the proposed matrix factorization algorithms. Section 5 shows an experimental evaluation and Section 6 presents some concluding remarks.

2 Related Work

Multimodal representations for images are usually related to the combination of two sources of information: visual features and text data [6]. The goal of combining these data sources is to complement the possible representations of image contents using semantic information extracted from text data. Previous works include probabilistic models [7] and matrix factorization algorithms [8]. These approaches require solving large optimization problems or computing expensive updating rules, which make them infeasible for large image collections. Our approach differs from the others, since the algorithm formulation can be gracefully scaled up to large image databases.

The strategy that we follow to scale up the matrix factorization algorithms is based on stochastic approximations. Mairal et al. [9] proposed an online setting for matrix factorization, specially designed for sparse coding. We follow similar ideas, with a simpler formulation that follows the stochastic gradient descent structure. Also, our main research focus is to generate multimodal image representations instead of sparse coding.

3 Image Representation

In this work, we consider the problem of image retrieval using as queries example pictures. So, users are expected to upload an image file to the retrieval system, and the system is expected to analyze its visual content to identify potentially relevant images. Those images selected by the system are presented to the user as results of his/her search. This image retrieval system requires a representation of the visual contents for each image. In this work, we follow the Bag-of-Features (BoF) approach to model image descriptors for search. Basically, this representation accounts for the occurrence of

visual patterns that can be seen in an image, with respect to a predefined dictionary or codebook. The final representation of images is a histogram that represents the visual structure of images.

Given the BoF descriptors for a database of images, we can build a matrix of visual patterns vs. images in the collection. Also, given the BoF descriptors for a query image, the system can compute the similarity of it with respect to images in the database using the histogram intersection.

4 Multimodal Indexing Using Matrix Factorization

A multimodal representation for images is proposed, with the goal of improving the response of a system that uses only visual data to search similar images. The approach used in this paper to build the multimodal representation is based on latent factors. A common latent space for visual and text data is learned, i.e., any of both data modalities can be projected from its original representation space to the common latent space. In this way, the resulting multimodal space to represent images incorporates semantic information together with visual contents, and so, can provide a better mechanism to match similar images.

The computational methods used in this work for learning such a multimodal space are based on matrix factorization. The proposed algorithm simultaneously decompose the matrices of visual and text data to find a low rank approximation of them, by solving an optimization problem. To this end, assume the availability of two matrices of data, one for visual features $V \in \mathbb{R}^{n \times l}$ and the second for text data $T \in \mathbb{R}^{m \times l}$. Both matrices have the same number of columns, corresponding to the number of images in the database: l. Let n be the number of visual features, i.e., the number of rows in the visual matrix, and let m be the number of text terms, i.e., the number of rows in the text matrix. The problem of multimodal decomposition is to find the matrices P, Q and H such that

$$V \approx PH$$
$$T \approx QH$$

where $H \in \mathbb{R}^{r \times l}$ is an encoding matrix for the multimodal latent representation of images, $P \in \mathbb{R}^{n \times r}$ and $Q \in \mathbb{R}^{m \times r}$ are the multimodal transformations for visual and text data respectively. The main idea behind this model is to find a common representation H for the visual and text data, which is known as the latent representation, together with the corresponding transformations from the latent space to the source data. The dimensionality of the latent space, r, is a fixed parameter, which indicates how many latent factors should be extracted from the data.

This simultaneous factorization of V and T can be found by solving an optimization problem that minimizes the following objective function:

$$\begin{aligned} \min_{P,Q,H} \quad & \frac{1}{2}\left(\|V - PH\|_F^2 + \|T - QH\|_F^2\right) \\ & + \frac{\lambda}{2}\left(\|P\|_F^2 + \|Q\|_F^2 + \|H\|_F^2\right) \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, and λ is a regularization parameter for the unknowns in the problem.

Gradient Descent Solution. The problem above has a non-convex objective function. However, the function is differentiable for all unknowns and the solution can be computed using gradient descent as follows:

$$P_{\tau+1} = P_\tau + \gamma\left(VH_\tau^T - P_\tau H_\tau H_\tau^T - \lambda P_\tau\right) \tag{2}$$

$$Q_{\tau+1} = Q_\tau + \gamma\left(TH_\tau^T - Q_\tau H_\tau H_\tau^T - \lambda Q_\tau\right) \tag{3}$$

$$H_{\tau+1} = H_\tau + \gamma\left(P_\tau^T V - P_\tau^T P_\tau H_\tau + Q_\tau^T T - Q_\tau^T Q_\tau^T H_\tau - \lambda H_\tau\right) \tag{4}$$

In the updating rules shown above, the subindex τ represents the solution at the iteration τ, and γ is the step size in the gradient descent algorithm. The solution above presents a batch formulation of the solution, i.e., at each step or iteration, the algorithm requires the full matrices V and T to decide the new direction of the solution. This can be quite expensive or even infeasible for large image collections, that can not be fit in memory. Alternative formulations of this problem run in parallel mode. However, our proposal is to formulate this problem using a stochastic gradient descent approximation.

Online Matrix Factorization. The idea of online learning using stochastic approximations is to compute the new solution for each unknown in the problem using only one data sample at a time [10]. Then, we can scan large data sets without memory restrictions, and this can be potentially scaled up to large image datasets. The stochastic gradient descent formulation for the multimodal matrix factorization problem is as follows:

$$h_\tau = \left(\lambda I + P_\tau^T P_\tau + Q_\tau^T Q_\tau\right)^{-1}\left(P_\tau^T v_\tau + Q_\tau^T t_\tau\right) \tag{5}$$

$$P_{\tau+1} = (1-\gamma\lambda)P_\tau + \gamma v_\tau h_\tau^T - \gamma P_\tau h_\tau h_\tau^T \tag{6}$$

$$Q_{\tau+1} = (1-\gamma\lambda)Q_\tau + \gamma t_\tau h_\tau^T - \gamma Q_\tau h_\tau h_\tau^T \tag{7}$$

where v_τ and t_τ are vectors of visual features and text features, respectively, for one image, and h_τ is the multimodal representation for that pair of vectors. This approach learns the matrices P and Q using one image with its corresponding text data at a time, and then, the image is discarded. For the image used in the iteration τ, the multimodal representation is approximated using the solution for h_τ, but this is also discarded with the image vectors v_τ and t_τ. Then, a final pass over the data would be required to recover the multimodal latent representation H for the full database, using the same expression for all images, and without updating P and Q.

Minibatch Extension. A minibatch extension of the algorithm presented in equations 6, 7 and 5 can be easily obtained by assuming that v_τ, t_τ and h_τ are not single vectors, but small matrices containing several vectors. The minibatch extension allows to process several images at the same time to make an update in the unknown parameters of the objective. Actually, the batch algorithm is an special case of the minibatch extension,

in which the number of images in the minibatch is equal to the number of images in the training set.

Indexing New Images. The algorithm and the solution presented are useful to learn the transformation matrices P and Q from a training image set that has visual and text data, and also to index a database by computing H using equation 5. For new images not included in the factorization analysis, we assume that all these matrices are already learned and are available. To index a new image, all what is required is the transformation matrices P and Q with a visual vector v for the new image. Since we expect queries with no text data, the following expression can be evaluated to project the new image to the multimodal latent space: $h_q = \left(\lambda I + P^T P + Q^T Q\right)^{-1} \left(P^T v\right)$. Other query strategies such as text queries and multimodal queries can also be supported in this framework, following similar extensions to those proposed in [8].

Searching in the Multimodal Space. After all images in the collection have been indexed, a new matrix with the latent representation that fuses visual features and text data is obtained: H. This matrix has as many columns as images in the database, so each image has a column vector $h \in \mathbb{R}^r$, with dimensionality r as the number of multimodal latent factors. Query images can be projected in the same space as well. So, to search in the multimodal latent space, we will use the dot product between these vectors, which accounts for the degree of similarity between two latent representations. Our assumption is that images with similar semantic interpretations will have similar multimodal factors in this representation.

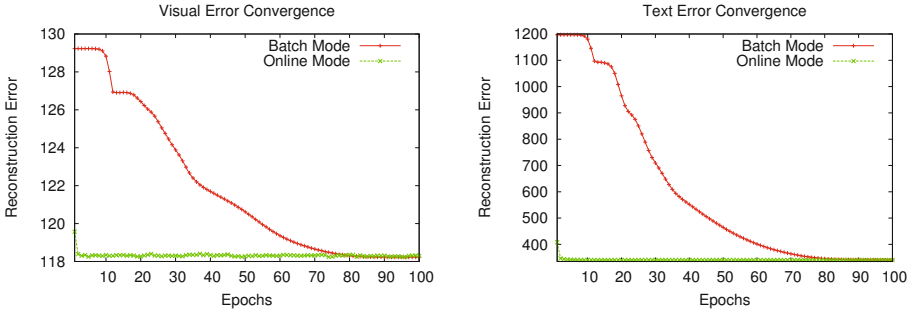
5 Experimental Evaluation

Data. The image collection used in this work is the Corel 5k image dataset, which is composed of 5,000 photographs organized in 50 categories, and has been used in different image retrieval evaluations by many researchers [7,11]. It consists of text annotations for each image, using a text dictionary with 374 terms. Also, the dataset has been organized in 3 parts to allow other researchers to reproduce experimental results: training (4,000 images), validation (500 images) and test (500 images). This data set is used to simulate retrieval performance using the 50 initial categories as ground truth. We follow the same experimental setup in this study. The visual matrix representation is built using a bag-of-features extracted with the same features as [8]: DCT coefficients in all color channels for local features and a dictionary of visual words with 2,000 clusters. The text matrix is built using a boolean vector representation: 1 for terms attached to images and 0 otherwise.

Algorithm Convergence. The first evaluation conducted in this work is the analysis of convergence of the algorithm using the batch and online approaches. The input matrices for training are $V \in \mathbb{R}^{2000\times 4000}$ and $T \in \mathbb{R}^{374\times 4000}$. We set the parameter $r = 50$, which defines the size of the multimodal latent space (or the number of latent factors). Then, the expected output matrices after the learning algorithm is run are $P \in \mathbb{R}^{2000\times 50}$, $Q \in \mathbb{R}^{374\times 50}$ and $H \in \mathbb{R}^{50\times 4000}$.

Both strategies, the batch and online algorithms, require the definition of the parameter λ for regularization, as well as the parameter γ, the gradient descent step size. We

set both parameters with the same values for both algorithms, using $\lambda = 0.1$ and defining γ as a decreasing function of time [10]: $\gamma(\tau) = \frac{\gamma_0}{1+\gamma_0\lambda\tau}$, where $\gamma_0 = 0.1$ is the initial parameter, and τ is the iteration number. Both algorithms were run during the same number of epochs, where an epoch is defined as a complete scan over the dataset.



(a) Visual Data (b) Text Data

Fig. 1. Evolution of the reconstruction error per epochs

Figure 1 present the findings of error convergence for 100 epochs, for both data modalities. These results show the evolution of the reconstruction error for the visual and text matrices for each epoch of the algorithm. The scale of the error units are specific for this data set. The results show that the online algorithm achieve the lowest reconstruction error from the very first epoch, converging much more faster than the batch algorithm. This means that the online algorithm provides a large reduction in the computational cost with respect to the batch counterpart, mainly because the online algorithm requires very few epochs to reach a stable solution. This tendency suggest that no more than 5 epochs are required to achieve a good factorization of the multimodal data.

To investigate how many iterations are actually needed to achieve a good performance, not in terms of reconstruction error, but in terms of retrieval accuracy, we run the algorithm for 20 epochs and evaluated the resulting model at each epoch by conducting information retrieval experiments. The experiments consist of using example images to search in the multimodal index, as is described in the next Subsection. The performance is measured using Mean Average Precision (MAP). In terms of MAP, the higher the value the better, since it suggests a more accurate response with respect to the results expected by users. The results show that the factorization achieves very good retrieval performance (a MAP of 0.2159), in contrast to the baseline based on image matching using visual contents only, which has a significantly lower performance (a MAP of 0.1239). More importantly, the improvement showed by the online multimodal indexing method is obtained with a single pass over the training data. Running the algorithm for more epochs does not seem to improve or hurt the quality of the factorization in a significant way. This is a very important result that supports the idea of scaling the algorithm to very large data sets with minimum processing effort: no more than a single scan on the available training data.

Retrieval Evaluation. To evaluate the retrieval performance, the validation data set is used as queries for the system. This allows to simulate a total number of 500 queries from 50 different categories. The queries in this system are received as image examples alone with no attached text or labels. However, the database has been indexed using multimodal data. Since queries do not have any text, it is reasonable to search using the BoF descriptor only to observe the benefit of using multimodal indexing strategies. Besides the proposed algorithm, we implemented and tested two different multimodal indexing algorithms recently proposed in the literature, which are based on Nonnegative Matrix Factorization (NMF): Multimodal NMF [12] and Asymmetric NMF [8].

Table 1 compares the retrieval performance of different indexing methods using MAP. The results show that using multimodal data for image search improves upon the baseline that uses only visual information, i.e., a multimodal indexing strategy for image retrieval can be used to deliver more meaningful results for users. The proposed online algorithm provides the second best retrieval performance in these experiments, with a very competitive result.

In addition, one of the most important aspects of the proposed algorithm is its ability to process the data set very quickly. For this experiment all the algorithms were implemented in Matlab and run in multithreading mode on a 12 core computer. The Table presents execution times and number of epochs required to achieve the reported retrieval performance for each method. The proposed online algorithm runs from 13 to 25 times faster than the other algorithms, even though an online approach does not fully exploit multithreading as the other algorithms do. This demonstrates the potential of our approach to process really large image databases.

Table 1. Retrieval performance for image search using Mean Average Precision (MAP). The learning time and the number of epochs to achieve that performance are also reported.

Method	MAP	Learning Time	Epochs
Visual Matching	0.1239	N.A.	N.A
Online Multimodal MF	0.2159	**2.3 secs**	**1**
Batch Multimodal MF	0.2115	58.6 secs	100
Asymmetric NMF [8]	0.2203	36.7 secs	100
Multimodal NMF [12]	0.2096	29.3 secs	100

6 Conclusions and Future Work

This paper has presented a matrix factorization algorithm for multimodal data analysis. The most remarkable characteristic of the proposed algorithm is its online formulation, which leads to very fast convergence over the batch algorithm. Experimental results show significant difference between the convergence speed of the algorithms, showing that only one pass through the data is enough to obtain state-of-the-art performance. As part of our future work, we plan to process bigger datasets with potentially hundreds of thousands of images to test the ability of the proposed algorithm to deal with large collections in terms of processing time as well as learning power.

References

1. Rasiwasia, N., Moreno, P.J., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Transactions on Multimedia 9(5), 923–938 (2007)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2), 1–60 (2008)
3. Fan, X., Xie, X., Li, Z., Li, M., Ma, W.: Photo-to-search: using multimodal queries to search the web from mobile devices. In: Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 143–150. ACM, Hilton (2005)
4. Muller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications: Clinical benefits and future directions. International Journal of Medical Informatics 73, 1–23 (2004)
5. Makadia, A., Pavlovic, V., Kumar, S.: A New Baseline for Image Annotation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 316–329. Springer, Heidelberg (2008)
6. Atrey, P., Hossain, M., El Saddik, A., Kankanhalli, M.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems (2010)
7. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary (chapter 7). In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
8. Caicedo, J.C., BenAbdallah, J., González, F.A., Nasraoui, O.: Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomput. 76, 50–60 (2012)
9. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res. 11, 19–60 (2010)
10. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of the 19th International Conference on Computational Statistics (2010)
11. Hare, J.S., Samangooei, S., Lewis, P.H., Nixon, M.S.: Semantic spaces revisited: investigating the performance of auto-annotation and semantic retrieval using semantic spaces. In: CIVR 2008: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, pp. 359–368. ACM, New York (2008)
12. Akata, Z., Thurau, C., Bauckhage, C.: Non-negative matrix factorization in multimodality data for segmentation and label prediction. In: 16th Computer Vision Winter Workshop (2011)