Supervised Models for Multimodal Image Retrieval based on Visual, Semantic and Geographic Information

Duc-Tien Dang-Nguyen, Giulia Boato, Alessandro Moschitti, Francesco G.B. De Natale
Department of Information and Computer Science - University of Trento
via Sommarive 5, 38123 Trento - Italy
{*dangnguyen, boato, moschitti*}*@disi.unitn.it, denatale@ing.unitn.it*

Abstract

Nowadays, large-scale networked social media need better search technologies to achieve suitable performance. Multimodal approaches are promising technologies to improve image ranking. This is particularly true when metadata are not completely reliable, which is a rather common case as far as user annotation, time and location are concerned. In this paper, we propose to properly combine visual information with additional multi-faceted information, to define a novel multimodal similarity measure. More specifically, we combine visual features, which strongly relate to the image content, with semantic information represented by manually annotated concepts, and geo tagging, very often available in the form of object/subject location. Furthermore, we propose a supervised machine learning approach, based on Support Vector Machines (SVMs), to automatically learn optimized weights to combine the above features. The resulting models is used as a ranking function to sort the results of a multimodal query.

1. Introduction

In the age of information technology, accessing and sharing digital media has become easier thanks to the widespread diffusion of the Internet as a global means for broadband communication. In particular, the availability of low-cost digital cameras along with the diffusion of social networks has favored the daily activity of capturing and sharing digital images. Unfortunately, this information is typically organized in an unstructured way, without following systematic rules. Therefore, different users (producers and consumers of the information) use inconsistent ways to define and use semantic tags for describing various images. This problem calls for new types of automatic search techniques able to deal with such uncertain information. In this perspective, multimodal indexing approaches can achieve better accuracy in image search applications thanks to their capability of jointly exploiting different sources of information (see for instance [14] [10] [16]). In previous work, this problem has been addressed in two different ways by using: (i) image-graph based techniques, or (ii) probabilistic models. In the first case, images are organized into graphs, where vertices represent images and edges measure the visual similarity among images. Then, a clustering method is applied to perform the retrieval. Both visual and semantic information are taken into account, creating a a bipartite graph in [4], [5], fusing them into a fusion graph in [9], or combining them from small tiles in [1]. Additionally, in [3] [7] geographic information is introduced as additional feature. In the second class of methods, probabilistic models are applied in different ways, e.g., using Combinatorial Markov Random Fields [2] or Probabilistic Latent Semantic Analysis (PLSA) [8].

In this paper, we propose a multimodal image analysis tool that properly combines visual, semantic and geographic information, following the second approach, and in particular exploiting a PLSA methodology. The underlying intuition is that (i) the visual features are strongly correlated with the image content; (ii) the user annotation (if reliable) gives hints about the main concepts conveyed by the image; and (iii) GPS coordinates (often available in recent photos) reliably provide geographic correlation. It should be noted that each of the above dimensions may be more or less effective depending on different queries or application domains, e.g., in case of professional photographer applications, where the annotation is more reliable and consistent, semantic tags play an important role. In contrast, in user generated contents, visual information can be more effective. Thus, we present a model based on SVMs [15], able to learn from the data the optimal weights to be assigned to multimodal descriptors. More in detail, we use data collected from users as follows:

1. we define a random set of image queries[1];

[1] They cannot be considered completely random, e.g., with respect to

2. we used our basic search engine to retrieve a set of images having the highest vector-based similarity with each query;

3. the images that are judged as relevant by human annotators are tagged as positive examples whereas the others are tagged as negative examples;

4. we train SVMs with such examples. The obtained model contains the weights for the visual, tag and GPS dimensions.

The description of a first method is presented in Section 2 whereas in Section 3 the SVM-based model and the resulting optimized multimodal approach is presented. Experimental results and discussion of both methods on a large database are given in Section 4.

2. Combining Visual, Concept and GPS signals

The proposed method takes into account three different types of information extracted from user generated multimedia contents: visual content, image tagging, and geo location. The three descriptors are then properly analyzed and combined into a unique multimodal similarity measure. In particular, PLSA is applied to both the visual and the tag feature spaces producing corresponding *topic* spaces with reduced dimensions, as in [13]. Indeed, PLSA enables the learning of an abstract high-level description, using occurrence counts of low-level features. The training phase is carried out using Expectation-Maximization. This allows for performing a very fast on-line retrieval also for very large datasets (see [6] for theoretic details).

Concerning visual features, we compute a Scale Invariant Feature Transform (SIFT) (i) by associating a 128 element descriptor with each salient point in the image; (ii) by defining a vocabulary with 2500 salient points using K-Means (training set of 5000 images); and (iii) by applying a *bag-of-words* model for associating a feature vector with each image. These vectors of word counts are used to compute a PLSA model with 100 topics, deriving a 100-dimensional description of each image, based on visual features.

To learn the PLSA model of image annotations, a vocabulary of tags is required. Our vocabulary consists of all the tags in the dataset, except words used just once or by a single user. The total number of terms selected is of 5500 words. Similarly, for the PLSA of visual features we use 100 topics, leading to a 100-dimensional description for each image, based on annotation.

The retrieval phase is performed by computing the nearest neighbors of the test image in the topic space, using L_1 or L_2 distance. Due to reduced dimensions of these spaces the procedure is very efficient. To perform the retrieval using visual (VS), and semantic topics, (TS), and also considering GPS coordinates (GPS), we propose the following metric:

$$Score_1 = GPS \times (\alpha \times VS + \beta \times TS), \qquad (1)$$

where α and β are parameters, which have to be manually set, and GPS, VS and TS are the scores computed according to L_1 or L_2. It should be pointed out that the idea of linearly combining multimodal descriptors has been already introduced in [11], although the use of multi-modality is further extended in our tool by introducing the geographic information to improve the final performance. The score concerning with location information is calculated as the distance between the GPS coordinates of the query and of the retrieved images. We tested the accuracy of this new feature and report the results in Section 4. In the experiments, we call the system using 1 "Multimodal 1" ($MM1$).

3. Supervised Multimodal Approach

Fine tuning of the parameters in Eq. 1 seems promising to significantly improve the retrieval accuracy. The standard approach relies on a development set (DS): for each query, it contains the relevant images, annotated by users. The impact of a parameter setting can be tested by computing the accuracy in retrieving relevant images (for our purpose we have only two labels: relevant/irrelevant). However, the optimization with respect to a DS can be tricky since: on one hand, we can surely fit the parameters for optimal retrieval accuracy. On the other hand, this may just overfit the parameters on DS whereas our image retrieval application should be useful for open domain search or at least for a large domain (e.g., the set of images referring to some broad topic).

Therefore, we are proposing to use SVMs to learn such parameters from annotated data. SVMs show two important properties: (i) they are robust to overfitting, offering the possibility to trade-off between generalization and empirical error, so we can also tune our model to a more general setting; (ii) we can include additional features in the parameter vector representation, which may be useful to adapt the value of the parameters for very different queries.

To exploit automatic tuning at the best, we prefer to use a slightly different version of Eq. 1, namely Eq. 2, illustrated below:

$$Score_2 = w_1 \times VS + w_2 \times TS + w_3 \times GPS, \qquad (2)$$

where w_i are the weights for the different dimensions and can be compactly represented by the vector $\vec{w}$.

To learn automatically such vector, we need training data T collected as described in the introduction. Let $\vec{x}_i$ be the

the Web, as they refer to a specific database.

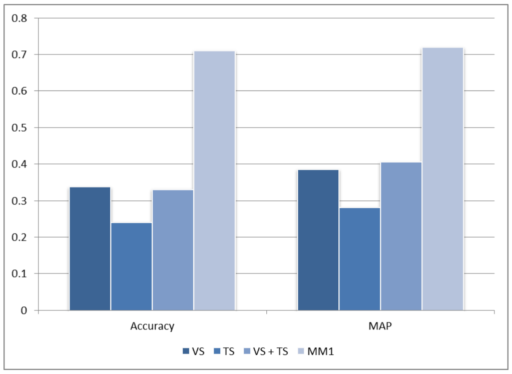

Figure 1. Performance comparison of $MM1$ and baseline in terms of accuracy and MAP.

vector containing the VS, TS and GPS values for each annotated image $x_i \in T$, we can apply the following optimization problem of SVMs:

$$\begin{array}{ll} min & ||\vec{w}|| + C\sum_{i=1}^{m} \xi_i^2 \\ s.t. & \begin{cases} y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1-\xi_i, & \forall i = 1,..,m \\ \xi_i \geq 0, \quad i=1,..,m, & \end{cases}\end{array} \tag{3}$$

where $\vec{w}$, ξ_i and b are learned parameters, m is the cardinality of T, y_i is the label assigned by the user to x_i and C is the trade-off parameter. The latter can be manually set to make our parameterization more adaptable to new domains (i.e., more general). We call this method “Multimodal 2” ($MM2$).

4. Experimental Results

The database used for the experimental evaluation consists of 100.000 images of Paris from Flickr. Visual vocabulary is composed by 2500 SIFT terms calculated on a training set of 50.000 images. Various vocabulary dimensions were tested but the results were not significant better (see also [12]). The tag vocabulary is defined by 5500 words as described above, using also in this case 50.000 images for training of PLSA.

Since the user annotations are sometimes the same for the whole album, in the retrieval phase we impose to consider maximum two images per user. This allows us to avoid considering many similar images taken by the same photographer.

The initial parameters of similarity metric of Eq. 1 were set to $\alpha = 0.8$ and $\beta = 0.2$ since the image retrieval based on visual features is almost always more reliable.

Methods	Accuracy %	MAP
VS	33.9	0.39
TS	24.2	0.28
$VS + TS$	34.2	0.40
$MM1$	71.1	0.72
$MM2$	72.0	0.78

Table 1. System comparisons according to accuracy and MAP: the two proposed multimodal approaches outperform methods exploiting only visual or/and semantic information.

We carried out experiments with 100 query images of Paris and we retrieved the top-ranked 9 images with our search engine. Such results were judge as relevant or irrelevant by 72 annotators. We finally consider the image relevant if more than half of the testing users considered it relevant and irrelevant otherwise.

Out of 900 retrieved images our system achieved the following results: (i) 305 relevant images only using the visual topic space (VS); (ii) 218 relevant images using the tag topic space (TS); (iii) 308 relevant images using the fusion of visual and tag information ($VS + TS$); and (iv) 641 relevant results using the first proposed multimodal approach $MM1$, which also exploits GPS coordinates (combines features). Figure 1 shows the comparison of both accuracy and Mean Average Precision (MAP) of the four different approaches mentioned above.

To compute the metric of Eq. 2 used in $MM2$, we trained SVMs with 70% of the images whereas we tested the models with the remaining 30%. The results show that $MM2$ achieves an accuracy of 72% and an MAP of 0.78, outperforming all the previous approaches. Table 1 shows a comparison of all the methods on the same test set above.

Figures 2 and 3 show the results for k retrieved images for each query, where the k values are listed on the x-axis. The graphs confirms that the proposed multimodal methods outperform the baseline for any k value.

Examples of some retrieved images according to different metrics, i.e., VS, TS, $VS + TS$, $MM1$ and $MM2$ are shown in Figures 4-8. It is worth noticing that: (1) the multimodal approaches improve the basic models especially in case the tag annotation is not reliable; and (2) the multimodal methods are promising to improve diversification of retrieval results, i.e., they reduce the number of images representing the same situation (e.g., night or day, perspective, point of view).

5. Conclusions

In this paper, we presented a novel way to combine visual information with tags and GPS information to improve the performance of image ranking in a large-scale database

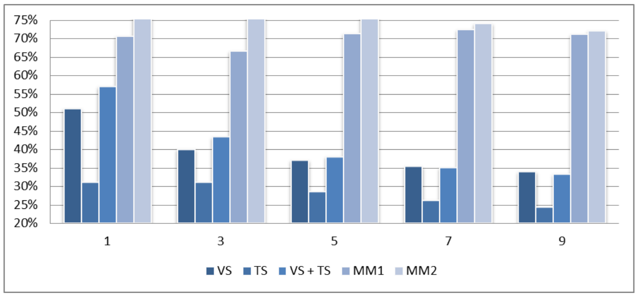

Figure 2. Systems' accuracy comparisons according to different k number of retrieved images for each query.

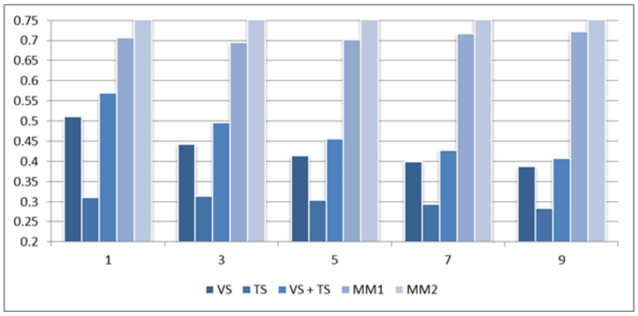

Figure 3. Systems' MAP comparisons according to different k number of retrieved images for each query.

image retrieval problem. Furthermore, we proposed a supervised machine learning approach, based on Support Vector Machines, to automatically learn the suitable weights for the features above. The experimental results confirm that the proposed approaches allow for improving accuracy of methods exploiting only visual information, only tags, and their combinations.

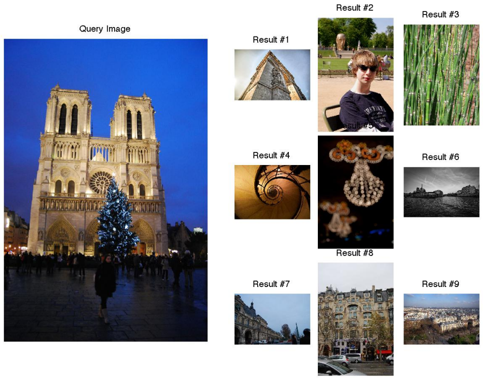

Figure 4. Example of retrieval based on tags. Only one result is confirmed as relevant due to non-reliable annotations.

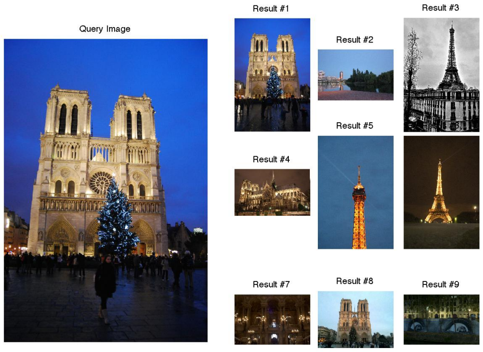

Figure 5. Example of retrieval based on visual features. Only two over nine images are relevant to the query.

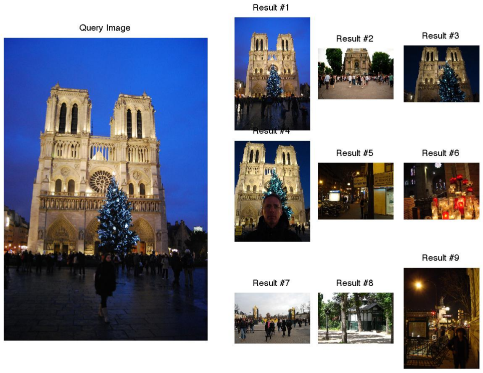

Figure 6. Example of retrieval based on the fusion of visual and tag features. Four over nine images are confirmed as relevant.

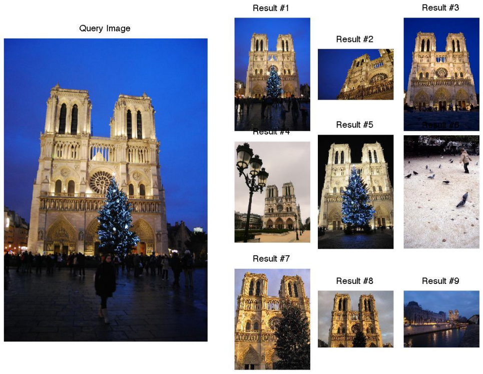

Figure 7. Example of retrieval following proposed MM1 approach. Eight over nine images are confirmed as relevant thanks to the support of GPS information.

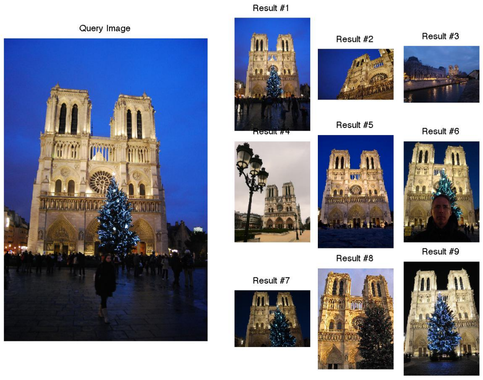

Figure 8. Example of retrieval following proposed MM2 approach. All results are confirmed as relevant.

6 Acknowledgments

This work has been partially supported by the EC project FP247758: Trustworthy Eternal Systems via Evolving Software, Data and Knowledge (EternalS).

References

[1] R. Agrawal, W. Grosky, and F. Fotouhi. Searching an appropriate template size for multimodal image clustering. *International Journal on Information and Communication Technologies*, 2(3-4):251–255, 2009.

[2] R. Bekkerman and J. Jeon. Multi-modal clustering for multimedia collections. In *IEEE Computer Vision and Pattern Recognition*, 2007.

[3] H. Frigui and J. Meredith. Image database categorization under spatial constraints using adaptive constrained clustering. In *IEEE International Conference on Fuzzy Systems*, pages 2268–2276, 2008.

[4] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *ACM International Conference on Multimedia*, 2005.

[5] B. Gao, T.-Y. Liu, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering. In *ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 41–50. ACM, 2005.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57. ACM, 1999.

[7] A. Jaffe, M. Naaman, T. Tassa, and M. Davis. Generating summaries and visualization for large collections of georeferenced photopraphs. *ACM International Workshop on Multimedia Information Retrieval*, 2006.

[8] R. Lienhart, S. Romberg, and E. Hörster. Multilayer pLSA for multimodal image retrieval. In *ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 9:1–9:8. ACM, 2009.

[9] S. Papadopoulos, C. Zigkolis, G. Tolias, Y. Kalantidis, P. Mylonas, Y. Kompatsiaris, and A. Vakali. Image clustering through community detection on hybrid image similarity graphs. In *IEEE International Conference of Image Processing*, 2010.

[10] R. Raguram and S. Lazebnik. Computing iconic summaries of general visual concepts. In *Computer Vision and Pattern Recognition Workshop*, 2008.

[11] F. Richter, S. Romberg, E. Hörster, and R. Lienhart. Multimodal ranking for image search on community databases. In *International conference on Multimedia information retrieval*, MIR '10, pages 63–72. ACM, 2010.

[12] C. Ries, S. Romberg, and R. Lienhart. Towards universal visual vocabulary. In *IEEE International Conference on Multimedia and Expo*, 2010.

[13] S. Romberg, E. Hoerster, and R. Lienhart. Multimodal plsa on visual features and tags. In *IEEE International Conference on Multimedia and Expo*, pages 414 – 417, 2009.

[14] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 33(4):754–766, 2011.

[15] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[16] G. Wang and D. A. Forsyth. Object image retrieval by exploiting online knowledge resources. In *IEEE Computer Vision and Pattern Recognition*, 2008.