Computational Models Enhancing Semantic Access to Digital Repositories

Floriana Esposito, Nicola Di Mauro, Claudio Taranto, and Stefano Ferilli

Department of Computer Science, University of Bari "Aldo Moro"
{esposito,ndm,claudio.taranto,ferilli}@di.uniba.it

Abstract. The growing amount of heterogeneous digital repositories has created a demand for effective and flexible techniques for automatic multimedia data retrieval. While the primary type of information available in documents is usually text, other type of information such as images play a very important role because they pictorially describe concepts that are dealt with in the document. Unfortunately, the semantic gap separating the visual content from the underlying meaning is wide.

The main goal of the project concerns the investigation of machine learning approaches to improve the semantic access to multimedia repositories by combining information gathered from the textual content with the one coming from pictorial representation. Furthermore, they have to be scalable, efficient and robust with respect to the inborn high-dimensionality and noise in the data collection.

1 Introduction and Motivation

The rapid expansion of digital data repositories raised problems that go beyond simple acquisition issues, and cause the need to organize and classify the contents in order to improve the effectiveness and efficiency of the retrieval procedure. During the past years a considerable effort was spent in the definition of automatic tools for features extraction from row and unstructured data, such as images, video, audio and text, resulting in the development of systems for content-based retrieval based on indexing and querying engines. However, they lack in dealing with one of the main characteristics of multimedia repositories represented by the existence of relations among the objects contained in the collection. Furthermore, as the volume and the typology of the data increases, memory and processing requirements need to correspondingly increase at the same rapid pace, and this is often prohibitively expensive. Indeed, multimedia digital collections on this scale make performing even the most common and simple indexing/retrieval task non trivial.

The National Projects "Progetto di Ateneo 2008 - Modelli computazionali con caratteristiche intelligenti per l'accesso semantico a documenti digitali" and "Progetto di Ateneo 2010 - Metodi e modelli per l'interpretazione semantica di immagini digitali" intended to investigate the applicability of relational models on complex data, such as multimedia digital collections, by proposing efficient

M. Agosti et al. (Eds.): IRCDL 2011, CCIS 249, pp. 107–110, 2011.

and robust methods to solve inference and learning tasks on noisy and high-dimensional data with the aim of providing meaningful annotations for making more effective the indexing and retrieval procedure.

2 Scientific Challenges

Information retrieval in large digital repositories is at the same time a hard and crucial task. This task is made more difficult by the presence of different object contents (image/video/audio/text).

One of the challenges of this research area is represented by the object description. Usually objects are described by low level features automatically extracted from raw data. The problem with them is that they seldom represent the semantic content of the object, which is commonly the focus of a user query. This phenomenon is known as the semantic gap between the object descriptor and the user search criteria. Thus the need of deriving high level semantic features to uniformly describe the objects in the collection. As a consequence, even the relationships among objects in the collection have to be taken into account and properly described and handled for completely outlining the repository.

On the other hand, the availability of massive, and consequently, noisy data have to be reshaped the design of indexing and querying approaches. To this aim, dimensionality reduction and statistical theory play pivotal roles.

3 Key Technologies

The development of an efficient and effective framework able to deal with multimedia collections must necessary take into account theoretical models able to take advantage of peculiarities and cope with problems specific of the domain. As to concern the first aspect, i.e. the peculiarities of the domain, it is unquestionable that a lot of relationships among data involved in a multimedia repository exist and that neglecting them can be dangerous for an effective management of the data. Furthermore, the problem of noise and high dimensionality of such kind of collections is a key issue in developing efficient indexing and querying engines. In our perspective, both the presented questions can be successful handled by exploiting relational learning paradigms.

Classical relational learning approaches fail in dealing with noisy data and mostly with high-dimensional ones. Hence, the need of investigating the applicability of probabilistic/statistic relational learning techniques to complex data. The representation and use of probability theory makes Statistical Relational Learning (SRL) techniques suitable for combining domain knowledge and data, expressing relationships, avoiding overfitting a model to training data, and learning from incomplete datasets. As for classical probabilistic graphical models, such as Bayesian networks and Markov networks, statistical relational languages exploit structure underlying many distributions one wants to encode. The same structure often allows the distribution to be used effectively for inference, answering queries using the distribution as a model of the world. Finally, this framework facilitates the effective construction of the models by learning from data a model that provides a good approximation to a past experience thus enhancing the retrieval step in noisy data collections.

4 Contribution by the Research Group

The contribution of the research unit focuses on the issues concerning the task of modelling and reasoning on relationships between images, or between objects within an image, by proposing efficient and robust to noise approaches.

In [1] the problem of complexity reduction for image indexing and retrieval was faced. Specifically, the aim of the work was twofold. It firstly investigated the possibility to efficiently extract an approximate distribution of the image features with a consequent indexing error reduction. Successively, the influence of such a resulting approximate distribution on the retrieval step performance based on similarity ranking was analysed. In particular, the image indexing process was improved by using a sampling method to approximate the distribution of correlograms, adopting a Monte Carlo approach to compute the distribution on a subset of pixels uniformly sampled from the original image. A further investigated variant was to sample the neighborhood of each pixel too.

Correlograms can be also used in a profitable way as features representing the images in a complex network of relations with the aim of improving image classification tasks. Specifically, in [2] a method to improve the classification accuracy adopting a Statistical Relational Learning approach is explored. The main idea is to assume that the images in a domain are not mutually independent and to try to elicit the hidden information representing the probabilistic connections between two images taking into account the possible relationships. To reach this goal, images are represented by means of a complex probabilistic network, where each image corresponds to a node and the connection degree between images is represented by a probabilistic edge. The relationship degree between images may be computed adopting a similarity measure based on their feature based representation. The ultimately goal was to verify whether modelling the problem of image classification using a SRL language can improve the accuracy of a classical K-Nearest Neighbour (KNN) approach. In particular, we adopted the probabilistic logic ProbLlog as SRL model to describe the structure of the probabilistic network arising from the abstraction process we adopted to represent an image collection.

In [3] it was proposed a method for Object Recognition that tries to understand an image by looking for known shapes in it, and relies on a combination of existing and novel image processing techniques, as a preliminary step to describe images using higher-level, human-understandable concepts and relationships among them. In particular, the approach focuses on the identification of potential objects in the image to be exploited in the indexing phase, on their representation and storage in suitable data structures and, lastly, on the definition of a retrieval algorithm that allows to detect known objects in new images.

Acknowledgements. The work was partially supported by the National Projects *"Progetto di Ateneo 2008 - Modelli computazionali con caratteristiche intelligenti per l'accesso semantico a documenti digitali"* and *"Progetto di Ateneo 2010 - Metodi e modelli per l'interpretazione semantica di immagini digitali"*.

References

1. Taranto, C., Di Mauro, N., Ferilli, S., Esposito, F.: Approximate image color correlograms. In: Bimbo, A.D., Chang, S.-F., Smeulders, A.W.M. (eds.) Proceedings of the 18th International Conference on Multimedia, pp. 1127–1130. ACM (2010)
2. Taranto, C., Di Mauro, N., Esposito, F.: Probabilistic Inference over Image Networks. In: Agosti, M., et al. (eds.) IRCDL 2011. CCIS, vol. 249, pp. 1–13. Springer, Heidelberg (2011)
3. Ferilli, S., Basile, T.M.A., Esposito, F., Biba, M.: A contour-based progressive technique for shape recognition. In: Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR), pp. 723–727. IEEE Computer Society, Washington, DC (2011)