Image Access and Data Mining: An Approach

Chabane Djeraba

IRIN, Ecole Polythechnique de l'Université de Nantes,
2 rue de la Houssinière, BP 92208 - 44322 Nantes Cedex 3, France
djeraba@irin.univ-nantes.fr

Abstract. In this paper, we propose an approach that discovers automatically visual relations in order to make more powerful the image access. The visual relationships are discovered automatically from images. They are statistical rules in the form of a → b which means: if the visual feature "a" is true in an image then the visual feature "b" is true in the same image with a precision value. The rules concern symbols that are extracted from image numerical features. The transformation of image numerical features into image symbolic features needs a visual feature book in which each book feature is the gravity center of similar features. The approach presents the clustering algorithm that creates the feature book.

1. Introduction

Image access associated to data mining may be seen as a new way of thinking retrieval of images and it opens up to a lot of new applications which have not been possible, previously. The new possibilities given by image access and data mining lies in the ability to perform "semantic queries-by-example", meaning that we can present an image of an object, color, pattern, texture, etc., and fetch the images in the image collections that most resemble the example of the query. For image collections the new possibilities lie in the ability to access efficiently and directly selected images of the database.

Our paper proposes a new way to extract automatically the hidden user semantics of images, based on basic content descriptions. Discovering hidden relations among basic features contributes to extract semantic descriptions useful to make accurate the image accesses. In our case, the relationship discovery are held into two important steps : symbolic clustering based on the new concept of visual feature book and relevant relationships discovery. The feature book is created on the basis of a new algorithm of global/local clustering and classification, powerful image descriptors and suitable similarity measures.

The rules extracted are composed of feature book items. In this paper, we consider textures and colors. The rules are qualified by conditional probability and implication intensity measures.

D.A. Zighed, J. Komorowski, and J. Zytkow (Eds.): PKDD 2000, LNAI 1910, pp. 375-380, 2000.

We organize the paper as follow : in section 2, we describe how the knowledge discovery is useful to content-based image retrieval. In section 3, we present how the relationships between image descriptors are extracted. In section 4, we describe some experiment results.

2. Semantic Queries-by-Example?

The central question is : how to extract and represent the content in order to obtain accurate image access ?

To obtain accurate image access, we consider semantic representations that include image class hierarchy (images of flowers, panorama, etc.) characterized by knowledge. «semantic query by examples» specifies a query that means «find images that are similar to those specified». The query may be composed of several images. Several images accurate the quality of retrieval. For example, Several images of a «waterfall» accurate the description of the waterfall. This property makes possible the refinement of retrieval based on the feed backs (results of previous queries).

In the retrieval task, features (colors, textures) of the query specification are matched with the knowledge associated to classes (ex. natural, people, industries, etc.). The suited classes are « Natural », then the matching process focus the search on the sub-classes of Natural : « Flowers », « Mountain », « Water », « Snow », etc. The knowledge associated to flowers and waterfalls are verified, so the matching process focuses the search on the « Flower » and « Water » classes. « Flowers » and « Water » classes are leaves, so the matching process compares the features of the examples with features of the image database to determine which images are similar to the example features. The matching task is based on computing the distance between target and source image regions. When mixing several features, such as colors and textures, the resulting distance is equal to the Sum taking into account the ponderation values of the considered features. The resulting images are sorted, the shortest distance corresponds to the most similar images.

An important advantage of the semantic queries by examples is the efficiency of the content-based retrieval. When the user gives examples of image to formulate his query, and asks "find images similar to the examples", the system will not match the source image with all the images in the database. It will match the source image features with only the target image features of suited classes. If the knowledge associated to a class is globally verified, then the considered class is the suited one. Then, the system will focus the search on the sub-classes of the current one. In the target classes that contain few instances, the search is limited to sequential accesses. Another advantage is the richness of descriptions contained in the results of queries since the system presents both similar images and their classes.

The semantic queries by example needs an advanced architecture. The advanced architecture supports the knowledge in the form of simple rules. Simple rules characterize each semantic class (flowers, natural, mountain, etc.), and are extracted automatically. The rules describe relationships between visual features (colors and textures of images). Each set of rules associated to a class summarizes image contents of the class. Rules contribute in the discrimination of each class, so they represent knowledge shared by the classes. When images are inserted in the database, it is classified "automatically" in the class hierarchy. At the end of the classification process, the image is inserted in a specific class. In this case, the distance between the image and the knowledge associated to the class is the shortest one, compared to the distance between the image and the other classes. Otherwise, the instantiation relationship between the image and the class, will not be considered.

This architecture avoids efficient retrievals and browsing through classes. For example, the user may ask "find images similar to the source image but only in People classes" or "find all images that illustrate the bird class with such colors and such shapes".

3. Discovery Hidden Relations

Based on image content description, the knowledge are discovered. The discovered knowledge characterizes visual properties shared by images of the same semantic classes (Birds, Animals, Aerospace, Cliffs, etc.).

The discovery is held into two steps : symbolic clustering and relationship discoveries and validation.

```
symbolic clustering
relationship discoveries and validation
```

In the first step, numerical descriptions of images are transformed into symbolic form. The similar features are clustered together in the same symbolic features. Clustering simplifies, significantly, the extraction process. For example, an image may be composed of region1 and region2. Region1 is characterized by light red color, and region2 by water color and water texture.

Light red color is not described by a simple string, but by a color histogram [Dje 00]. Even if the region colors of different images of the same class are similar (i.e. light red), the histograms (numerical representation of color) associated with them are not generally identical.

In the second step, the knowledge discovery engine determines automatically common features between the considered images in rule form. These rules are relationships in the form of `Premise => Conclusion` with a certain accuracy. These rules are called statistical as they accept counter-examples.

```
(texture, water) => (color, heavy_light) (P.C. 100 %, I.I
96.08 %), (texture, waterfall) => (color, white) (P.C.
100 %, 87.4327 %), (texture, texture_bird) => (color,
color_bird) (100 %, 40,45 %)
```

We implemented a technique that clusters numerical representation of color, texture [Dje 00], by using data quantization of colors and textures, we use also the term of feature book creation. The color and texture clustering algorithms are similar, the difference is situated in the distance used.

3.1 Principle of the Algorithm

The algorithm is a classification approach based on the following observation. The scalar quantification of Lloyd developed in 1957 is valid for our vectors (color histogram, Fourier coefficients [Dje 97]), four rate distribution and for a large variety of distortion criteria. It generalizes the algorithm by modifying the feature book iteratively. This generalization is known by k-means [Lin 80]. The objective of the algorithm is to create a feature book, based on automatic classifications themselves based on a learning set. The learning set is composed of feature vectors of unknown probability density. Two steps should be distinguished :

- A first step of classification that clusters each vector of the learning set around the initial feature book that is the most similar. The objective is to create the most representative partition of the vector space.
- A second step of optimization that permits the correct adaptation in a class of the feature book vector. The gravity center of the class created in the previous step is computed.

The algorithm is reiterated in the new feature book in order to obtain a new partition. The algorithm converges to stable position by evolving at each iteration the distortion criteria. Each application of the iteration of the algorithm should reduce the mean distortion. The choice of the initial feature book will influence the local minimum that the algorithm will achieve, the global minimum corresponds to the initial feature book. The creation of the initial feature book is inspired of the splitting technique [Gra 84].

The splitting method decomposes a feature book Y_k into two different feature books $Y_{k-\varepsilon}$ and $Y_{k+\varepsilon}$, where ε is a random vector of weak energy, and its distortion depends of the distortion of the splited vector. The algorithm is then applied on the new feature book in order to optimize the reproduction vectors.

3.2 Algorithm

Based on the learning set of length equal to T, the algorithm finds a feature book of colors and textures of length equal to L, that are the most representative colors and textures of image databases.

Global Clustering

```
FeatureBook $Y_f$ = SymbolicClustering (visual feature = VisualFeature, learning set = LearningSet, $Y_0$, T, L)
```

Local clustering

```
FeatureBook $Y_f$ = Clustering(visual feature = VF, learning set = LS, $Y_0$, $Y_f$, T, L, E)
```

The experimental results showed that the distortion values decrease quickly compared to splitting evolution. After the quick decreasing, the distortion values decrease very slowly. Conversely, The entropy increase quickly compared to splitting evolution, and then, it increases very slowly.

3.3 Relationship Discoveries and Validation

Based on the feature book, the discovery engine is triggered to discover the shared knowledge in the form of rules, and this constitutes the second step of the algorithm.

Accuracy is very important in order to estimate the quality of the rules induced. The user should indicate the threshold above which rules discovered will be kept (relevant rules). In fact, the weak rules are rules that are not representative of the shared knowledge. In order to estimate the accuracy of rules, we implement two statistical measures : conditional probability and implication intensity. The conditional probability formula of the rule a => b makes it possible to answer the following question: ''what are the chances of proposition b being true when proposition a is true ? Conditional probability allows the system to determine the discriminating characteristics of considered images. Furthermore, we completed it by the intensity of implication [Gra 82]. For example, implication intensity requires a certain number of examples or counter-examples. When the doubt area is reached, the intensity value increases or decreases rapidly contrary to the conditional probability that is linear. In fact, implication intensity simulates human behavior better than other statistical measures and particularly conditional probability. Moreover, implication intensity increases with the considered population sample representativity. The considered sample must be large enough in order to draw relevant conclusions. Finally, implication intensity takes into consideration the sizes of sets and consequently their influence.

4. Conclusion

To demonstrate the efficiency of the semantic content-based queries, the results of the semantic content-based queries are compared with the results of queries that do not use semantic content-based queries. Since it is not possible to retrieve all relevant images, our experiment evaluates only the first 50 ranked images.

Judging on the results, it is obvious that the use of knowledge leads to improvements in both precision and recall over majority queries tested. The average improvements of advanced content-based queries over classic content-based queries are 23% for precision and 17 % for recall. Precision and recall are better for semantic-based queries than for queries that use only visual features such as color and textures.

References

[Dje 00] Djeraba C., Bouet M., Henri B., Khenchaf A. « Visual and Textual content based indexing and retrieval », to appear in International Journal on Digital Libraries, Springer-Verlag 2000.

[Gra 82] Gras Régis, THE EISCAT CORRELATOR, EISCAT technical note, Kiiruna 1982, EISCAT Report 82/34, 1982.

[Gra 84] Gray R. M. « Vector Quantization », IEEE ASSP Mag., pages 4-29, April 1984.

[Gup 97]Amarnath Gupta, Ramesh Jain «Visual Information Retrieval», A communication of the ACM, May 1997/Vol. 40, N°5.

[Haf 95] Hafner J., al. «Efficient Color Histogram Indexing for Quadratic Distance Functions». In IEEE Transaction on Pattern analysis and Machine Intelligence, July 1995.

[Jai 98] Ramesh Jain: Content-based Multimedia Information Management. ICDE 1998: 252-253

[Lin 80] Linde Y., Buzo A., Gray R. M. « An algorithm for Vector Quantizer Design », IEEE Trans. On Comm., Vol. COM-28, N° 1, pages 84-95, January, 1980.

[Moo 51]Moores C. N. «Datacoding applied to mechanical organization of knowledge» AM. Doc. 2 (1951), 20-32.

[Rag 89] Raghavan, V., Jung, G., and Bollman, P., "A Critical Investigation of Recall and Precision as Measures", ACM Transactions on Information Systems 7(3), page 205-229

[Rij 79] C. J. Keith van Rijsbergen «Information retrieval», Second edition, London: Butterworths, 1979

[Sal 68] Salton Gerard «Automatic Information Organization and Retrieval», McGraw Hill Book Co, New York, 1968, Chapter 4.

[Zah 72] C. T. Zahn, R. Z. Roskies, « Fourier descriptors for plane closed curves », IEEE Trans. On Computers, 1972.