Image retrieval in the presence of important scale changes and with automatically constructed models

Cordelia Schmid, Krystian Mikolajczyk

IMAGE RETRIEVAL
in the presence of important scale changes and with automatically constructed models

Cordelia Schmid and Krystian Mikolajczyk

INRIA Rhône-Alpes GRAVIR-CNRS
655 av. de l'Europe, 38330 Montbonnot, France
Cordelia.Schmid@inrialpes.fr

Abstract

In this paper we address two aspects of image retrieval. First, we present the retrieval of an object or a scene in the presence of important scale changes. The approach is based on the detection of scale invariant interest points. These points are used to characterize the image ; the scale associated with each point allows to compute scale invariant descriptors. Our descriptors are, in addition, invariant to image rotation, to affine illumination changes and robust to limited perspective deformations. Experimental results for retrieval show an excellent performance up to a scale factor of 4 for a database with more than 5000 images.

Secondly, we automatically construct visual models for the retrieval of similar images. Models are constructed from a set of positive and negative sample images where no manual extraction of significant objects or features is required. Our model allows to efficiently capture "texture-like" structure and is based on two layers : "generic" descriptors and statistical spatial constraints. The selection of distinctive structure increases the performance of the model. Experimental results show a very good performance for retrieval as well as localization.

1. Introduction

The growing number of images has increased the need for tools which automatically search image collections. While tools based on keywords exist, they have two major drawbacks. Firstly, each image in the collection has to be described by keywords which is extremely time consuming. Secondly, the expressive power of keywords is limited and cannot be exhaustive. Consequently, a significant need for image content based tools exists. Existing tools can be categorized into those that search for specific objects and those that search for generic objects or similar images.

The difficulty in object indexing is to determine the identity of an object under arbitrary viewing conditions in the presence of cluttered real-world scenes or occlusions. Local characterization has shown to be well adapted to this problem. The small size of the characteristic regions makes them robust against occlusion and background changes. To obtain robustness to changes of viewing conditions they should also be invariant to image transformations. Recent methods for indexing differ in the type of invariants used. Rotation invariants have been presented by [8], rotation and scale invariants by [3] and affine invariants by [10]. All of these methods are limited to a scale factor of 2. The method presented in section 2 allows to deal with more important scale changes.

The difficulty in retrieving similar image is the definition of similarity which should be meaningful to the user. The first image retrieval systems were based on the comparison of global signatures, such as colour or texture histograms [5]. Results of these systems have shown to be unsatisfactory, as they do not represent the "semantic" image content ; they do not allow to find images containing instances of a model, as for example faces or zebras. More recent methods construct models and localize them in the image. They differ in the model representation and in the learning algorithm. Models are for example represented by global images patches [9], geometric relations of parts [11] or statistical models [6]. Learning algorithms are either supervised or unsupervised. Supervised algorithms require the manual extraction of regions or features. In the unsupervised case images are labelled as positive or negative which avoids time consuming manual intervention. The approach presented in section 3 is unsupervised and is based on a novel probabilistic model representation. It allows to learn a flexible statistical model which efficiently captures visual structure common to the positive and rare in the negative examples.

2. Retrieval based on scale invariant interest points

In this section we present an approach which allows indexing in the presence of scale changes up to a factor of 4. For more details the reader is referred to [4]. The success of this

This work was supported by the French project RNRT AGIR and the European FET-open project VIBES.

Figure 1: Image retrieval in the presence of image rotation and a scale factor of 4.9. The image database contains more than 5000 images. On the left the query image and on the right a few images of the database. The corresponding image is correctly retrieved (second image on the bottom row).

method is based on a repeatable and discriminant point detector. This detector is based on two results on scale space: 1) Interest points can be adapted to scale and give repeatable results [1]. 2) Local extrema over scale of normalized derivatives indicate the presence of characteristic local structures [2]. The first step of our approach is to compute interest points at several scale levels. We then select points at which a local measure (the Laplacian) is maximal over scales. This allows to select a subset of the points computed in scale space. For these points we know their scale of computation, that is their characteristic scale. Points are invariant to scale, rotation and translation as well as robust to illumination changes and limited changes of viewpoint.

We characterize an image by a set of scale invariant interest points. At each point we compute a descriptor at its characteristic scale. Descriptors are based on Gaussian derivatives and are invariant to image rotation and affine illumination changes. A voting algorithm is used to select the most similar images in the database. For each point of a query image, its descriptor is compared to the descriptors in the database using the Mahalanobis distance. If the distance is less than a threshold, a vote is added to the corresponding database image.

An example for retrieval is presented in figure 1. The database contains more than 5000 images; the images are extracted from 16 hours of video sequences which include movies, sport events and news reports. Similar images are excluded by taking one image per 300 frames. The total number of descriptors in our database is 2539342. The image which corresponds to the query image on the left is retrieved correctly (second image on the bottom row). In figure 2 we show the interest points, the initial matches and the matches after estimation of the homography for the example of figure 1.

(a) Extracted interest points

(b) Initial points matches

(c) Inliers to the estimated homography

Figure 2: Matches obtained for figure 1. (a) There are 190 and 213 points detected in the left and right images. (b) 58 points are initially matched. (c) There are 32 inliers to the estimated homography, all of which are correct. The estimated scale factor is 4.9 and the estimated rotation angle is 19 degrees.

3. Retrieval with automatically constructed models

In this section we present the automatic construction of models which efficiently capture visual structure common to the positive and rare in the negative examples. For more details the reader is referred to [7]. The visual structure is represented by "generic" descriptors and the joint probability of their frequencies over neighbourhoods. It can represent textures, for example the stripes of a zebra, as well as highly structured patterns, for example faces. The "generic" descriptors as well as the spatial frequencies are rotationally invariant. This allows to group similar but rotated patterns, as for example horizontal and vertical stripes of a zebra. It also makes the method robust to model deformations, as for example in the case of a cheetah sitting instead of standing upright. The rotational invariance as well as the flexibility of our constraints (spatial-frequency constraints instead of geometric constraints) permit our model to handle deformable objects, for example "textured" animals. Geometric constraints are useful for modelling object classes with similar spatial structure, for example faces, but do not allow to model deformable objects (animals, humans, etc.).

The steps of our model construction are the following. We first compute local rotationally invariant "Gabor-like" feature vectors at each pixel location. A clustering algorithm extracts "generic" descriptors for the collection of positive and negative images. The "generic" descriptors represent groups of similar feature vectors which occur if structure is repeated in the image or between images. The next step is to estimate the joint probability of their frequencies over neighbourhoods. These probabilities are multi-modal and are represented by a set of "spatial-frequency" clusters. Each cluster captures visual similar patterns. We do not estimate the global joint probability, but the conditional joint probabilities with respect to the "generic" descriptor at the center location. This allows to verify the coherence of the neighbourhood with respect to the center and adds a supplementary constraint ; the addition of conditional probabilities has shown to increase performance. The selection of distinctive "spatial-frequency" clusters determines characteristic model structure (common to the positive and rare in the negative examples). It allows to eliminate background patterns and to keep distinctive patterns of the model.

For our experimental results we constructed models from 15 sample images (5 positive and 10 negative). Our database contains 600 images of the corel dataset and 60 face images. We have learnt and tested 4 different models : a zebra model, a cheetah model, a giraffe model and a face model. Our database contains approximatively 60 images of each category, 5 of which are part of the training set and excluded from the test set. Equivalently, negative examples of the training set are not included in the test set.

The top row of figure 4 shows a subset of the training images (3 positive and 2 negative examples) used to learn the zebra class. The remaining rows display the 15 retrieved images with the highest probability score for the zebra class. The images are ordered by their probability score (from left to right and from top to bottom). The 14 most similar images are zebras; the 15th image is incorrectly retrieved. This incorrect retrieval is due to high probabilities for the branches which are visually similar to zebra stripes. It could be easily eliminated by adding a global constraint.

Our method also allows to localize the model in a retrieved test image by selecting locations with a high probability score. Results of localizing the zebra model are presented in figure 3. The locations with high scores are displayed in black. The body of the animal and three of its legs are correctly detected. Comparable results for localization of animals have to our knowledge not been presented before.

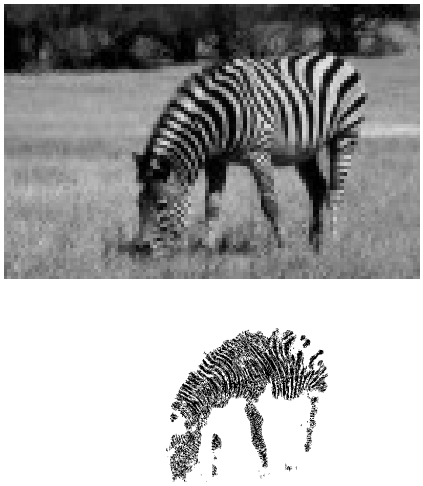

Figure 3: Localization of the zebra model for one of the test images (top). Locations with the high probability scores are displayed in black (bottom).

4. Conclusion and Discussion

In the first part of this paper, we have presented an algorithm for indexing that is invariant to important scale changes ; results are excellent up to a scale factor of 4. Furthermore, our approach is invariant to image rotation and translation as well as robust to illumination changes and limited changes in viewpoint. Performance could be further improved by using more robust point descriptors. In our future research, we intend to focus on the problem of affine invariance of point descriptors.

Secondly, we have presented a novel approach for model construction which significantly improves on the state of the art. Our model representation allows to capture efficiently

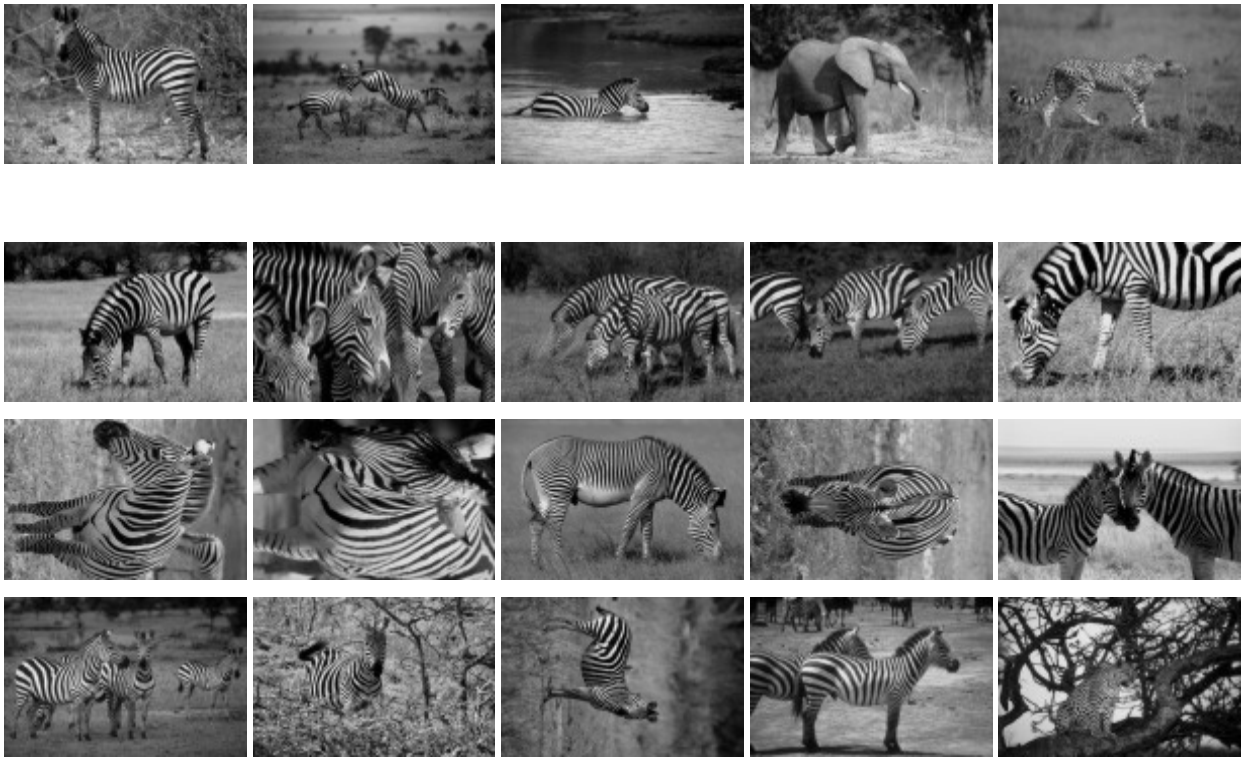

Figure 4: Retrieval results. The top row shows a subset of the training images (3 positive and 2 negative examples). The other rows show the first 15 retrieved images ordered by their score (from left to right and from top to bottom).

"texture-like" visual structure ; our learning algorithm is unsupervised and therefore does not require manual extraction of objects or features. It allows to learn an appropriate representation of the model. We are currently investigating five extensions. The first is to add scale selection to the model construction. The second is to learn which components of our multi-valued descriptors are significant. The third is to improve the clustering algorithm and to automatically select the number of clusters. The fourth is to include global constraints, for example by modelling relations between parts. The fifth extension is to improve the model over time by user interaction.

5. References

[1] Y. Dufournaud, C. Schmid, and R. Horaud. Matching images with different resolutions. In *CVPR*, pages 612–618, 2000.

[2] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.

[3] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[4] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, pages 525–531, 2001.

[5] W. Niblack, R. Barber, W. Equitz, M. Fickner, E. Glasman, D. Petkovic, and P. Yanker. The QBIC project: Querying images by content using color, texture and shape. In SPIE *Conference on Geometric Methods in Computer Vision II*, 1993.

[6] T.D. Rikert, M.J. Jones, and P. Viola. A cluster-based statistical model for object detection. In *ICCV*, pages 1046–1053, 1999.

[7] C. Schmid. Constructing models for content-based image retrieval. In *CVPR*, 2001.

[8] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–534, 1997.

[9] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *PAMI*, 20(1):39–51, 1998.

[10] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Visual99*, pages 493–500, 1999.

[11] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV*, pages 18–32, 2000.