Constructing models for content-based image retrieval

Cordelia Schmid

To cite this version:

Cordelia Schmid. Constructing models for content-based image retrieval. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '01), Dec 2001, Kauai, United States. pp.11–39, 10.1109/CVPR.2001.990922 . inria-00548274

HAL Id: inria-00548274

https://hal.inria.fr/inria-00548274

Submitted on 21 Dec 2010

Constructing models for content-based image retrieval

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

Abstract

This paper presents a new method for constructing models from a set of positive and negative sample images ; the method requires no manual extraction of significant objects or features. Our model representation is based on two layers. The first one consists of "generic" descriptors which represent sets of similar rotational invariant feature vectors. Rotation invariance allows to group similar, but rotated patterns and makes the method robust to model deformations. The second layer is the joint probability on the frequencies of the "generic" descriptors over neighborhoods. This probability is multi-modal and is represented by a set of "spatial-frequency" clusters. It adds a statistical spatial constraint which is rotationally invariant. Our two-layer representation is novel ; it allows to efficiently capture "texture-like" visual structure. The selection of distinctive structure determines characteristic model features (common to the positive and rare in the negative examples) and increases the performance of the model. Models are retrieved and localized using a probabilistic score. Experimental results for "textured" animals and faces 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, for example in stock photo agencies.

The first image retrieval systems were based on the comparison of global signatures, such as color or texture histograms [11]. 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 [17], geometric relations of parts [18] or statistical models [14]. Learning algorithms are either supervised or unsupervised. Supervised algorithms require the manual extraction of regions or features. In the unsupervised case images are labeled as positive or negative which avoids time consuming manual intervention.

In this paper we propose an unsupervised approach which constructs a model from a collection of positive and negative images. We introduce 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. The visual structure is represented by "generic" descriptors and the joint probability of their frequencies over neighborhoods. 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 modeling 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 neighborhoods. These probabilities are multimodal 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 neighborhood 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.

Related work Sung and Poggio [17] model a face as a rigid global patch and the distribution of these patches is learnt from a large collection of face images. Face patches have to be segmented manually. Schneiderman and Kanade [16] model faces and cars as sets of attributes (local histograms of wavelet coefficients). Positions of each attribute are represented with respect to a coordinate frame fixed to the object. Their representation is rigid, but allows for small positional variations. It is learnt from a large set of manually extracted examples. Characteristic distributions of feature vectors are learnt by [8, 14]. These methods require the manual extraction and annotation of regions. They can represent for example faces, road or sky. Amit and Geman [1] learn a hierarchical model from edge features. They select distinctive local feature groupings of edgels constrained by loose geometrical relationships and then build global spatial arrangements. Their method assumes that training images (faces) are registered with respect to a reference grid. Weber et al. [18] use a flexible shape model of distinctive rigid parts. The variability within a class is represented by a joint probability on the shape and part detectors. Their learning algorithm does not require the extraction of faces, but only assumes that the example images are labeled as positive or negative. Amit and Geman [1] as well as Weber et al. [18] learn the model representation, but their model is based on geometric shape and is therefore limited to spatially similar objects, as for example faces and cars. Ratan et al. [13] also learn visual concepts from a collection of positive and negative examples. Their system uses segmented regions as well as detected circles as initial description. It therefore depends on these results and the choice of a circle detector is object specific (adapted to the car class).

Overview This paper is organized as follows. Section 2 presents the extraction of “generic” descriptors. The construction of “spatial-frequency” clusters is presented in section 3. In section 4 the significance of each “spatial-frequency” cluster is determined. The probabilistic score for retrieval and localization of model instances is explained in section 5. Results are shown in section 6.

2. Generic descriptors

We represent local greyvalue structure by rotationally invariant feature vectors which are computed at each pixel location. These multi-dimensional feature vectors are in the following referred to as greyvalue descriptors. Greyvalue structure can be repeated in the image (in the case of texture) or between images (similar visual structure) ; it can also have similar values in a region. To summarize the information it is therefore appropriate to form groups of similar descriptors and describe them by their mean and variance. These groups are obtained by clustering multi-dimensional feature vectors and are in the following referred to as “generic” descriptors. Similar descriptors have been proposed previously. Rikert et al. [14] use a simple clustering algorithm to extract clusters of similar descriptors from a large set of sample images and then select significant clusters. Malik et al. [10] use the k-means algorithm to cluster descriptors of one image. They call the centers textons and uses them for a compact texture representation.

2.1. Greyvalue descriptors

Our greyvalue descriptors $\mathbf{d}_l$ are computed for each image pixel location $\mathbf{p}_l$. These descriptors are rotationally invariant and are obtained by convolution with isotropic “Gabor-like” filters. These filters combine frequency and scale :

$$F(x,y,\tau,\sigma) = F_0(\tau,\sigma) + \cos\left(\frac{\sqrt{x^2+y^2}\,\pi\,\tau}{\sigma}\right)\, e^{-\frac{x^2+y^2}{2\,\sigma^2}}$$

where τ is the number of cycles of the harmonic function within the Gaussian envelope of the filter, commonly used in the context of Gabor filters [5]. $F_0(\tau,\sigma)$ is added to obtain a zero DC component. This makes the filters robust to illumination changes, as we obtain invariance to intensity translations.

For our experiments we use 13 filters with scales σ between 2 and 10 and τ between 1 and 4. For smaller scales only small τ are used to avoid high frequency responses. Compared to [6, 14, 15] who use 24 filters or more, our description is of lower dimensionality. This avoids problems inherent to high-dimensionality of descriptors. Our results show robustness to limited scale changes. Scale invariance can be obtained by using scale selection [9] to determine the appropriate scale for computation. This is currently under investigation. A comparison of our “Gabor-like filters” with rotational invariant combinations of derivatives [7] has shown that the “Gabor-like” filters improve the performance.

2.2. Extraction of “generic” descriptors

Our “generic” descriptors are groups of similar greyvalue descriptors. These groups are obtained by clustering with a k-means algorithm [3]. We extract "generic" descriptors for the set of positive and negative sample images. Negative images are included to obtain a more descriptive set of "generic" descriptors. Negative "generic" descriptors permit to eliminate non-model descriptors. This avoids rejection based on a threshold.

The k-means algorithm finds k centers such that after assigning each data vector to the nearest center, the sum of the squared distance from the centers is minimized. Note that the k-means algorithm will only achieve a local minimum of this criterion. Our algorithm first normalizes the descriptors $\mathbf{d}_l$ using their mean and variance to avoid scaling effects. We then iteratively choose k centers such that after assigning each data vector to the nearest center, the sum of the squared distance from the centers decreases. Once the algorithm has converged to k clusters with centers $\boldsymbol{\mu}_i$, the covariance matrix of each cluster is computed from the descriptors assigned to it. Our k clusters are described by $C_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

The choice of the optimal number of clusters k is difficult and depends on the context. In the context of region segmentation, a small number of clusters is required, for example in [2] the number varies between 2 and 5. Such clusters have an important variance and are not appropriate in our context, as they are not sufficiently distinctive. A more important number k of clusters is therefore required [10, 14].

Figure 1 shows three "generic" descriptors (clusters) computed for the cheetah image on the left. The two leftmost cluster images display "generic" descriptors which characterize the cheetah, the cluster on the right represents background. In this figure a cluster is represented by the image locations at which the descriptors have the highest probability for this cluster.

2.3. Probability of a "generic" descriptor

We now define the probability of a "generic" descriptor C_i. For a pixel location $\mathbf{p}_l$ or equivalently for its greyvalue descriptor $\mathbf{d}_l$, the probability $P(C_i|\mathbf{d}_l)$ is defined by :

$$P(C_i|\mathbf{d}_l) = \frac{P(\mathbf{d}_l|C_i)P(C_i)}{P(\mathbf{d}_l)} = \frac{P(\mathbf{d}_l|C_i)P(C_i)}{\sum_{i=1}^{k} P(\mathbf{d}_l|C_i)P(C_i)} \quad (1)$$

We assume in the following that the clusters C_i are equally probable. The probability $P(\mathbf{d}_l|C_i)$ is computed by approximating the distribution of a "generic" descriptor $C_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with a Gaussian. We can then select for each image location $\mathbf{p}_l$ the most probable "generic" descriptor, that is the one with the maximum probability $P(C_i|\mathbf{d}_l)$:

$$C^*(\mathbf{p}_l) = C^*(\mathbf{d}_l) = \operatorname*{argmax}_{C_i} p(C_i|\mathbf{d}_l) \quad (2)$$

The most probable cluster is stored in a label image at the corresponding pixel location. Labels vary from 1 to k with k the number of clusters.

3. "Spatial-frequency" clusters

A second layer of information increases the distinctiveness of our representation. This layer is based on the "spatial-frequency" clusters which are more distinctive than simple "generic" descriptors. They allow an additional verification, that is permit to reject descriptors which accidently correspond to "generic" descriptors of the model. Our "spatial-frequency" clusters represent the joint probability on the frequencies of "generic" descriptors over a neighborhood. This probability is multi-modal ; our experiments have shown that it is clearly not sufficient to describe the distribution by its mean and variance. We do not estimate the global joint probability, but the conditional joint probabilities with respect to the descriptor of the center location. This allows to verify the coherence of the neighbors with respect to the center and has shown to add a supplementary constraint. We ignore the geometric spatial relationship of the "generic" descriptors, as we only use their frequencies. Note that frequencies are rotationally invariant.

Most of the spatial constraints proposed previously are based on geometric shape information [1, 18]. Geometric shape constraints allow to represent object classes which share features that are visually similar and occur in similar spatial configurations. Examples for such classes are faces or cars. Such constraints are not adapted for "textured" deformable objects such as animals, as they do not have similar spatial structure. The geometric structure of a cheetah for example is very different, if it is sitting or standing upright.

Distributions of descriptors over neighborhoods have been previously used by Schneiderman and Kanade [16]. They use attribute histograms over neighborhoods. Neighborhoods are fixed with respect to a reference frame, that is the local distributions have to occur in similar spatial positions. Their model is therefore not adapted to deformable objects. Furthermore, they do not learn the joint probability from examples and do not select the distinctive parts of the distribution. In the context of image segmentation, Malik et al. [10] compare windowed texton histograms, where the windows are centered around the two pixels being compared. This allows to decide on the presence of a region boundary. They do not attempt to learn a model.

3.1. Extraction of spatial-frequency clusters

In section 2.3 we have introduced a label image. Each label represents the most probable "generic" descriptor for the greyvalue descriptor computed at the image location. The label image is used to compute for each image location the frequencies (probabilities) of the "generic" descriptors C_i over a neighborhood :

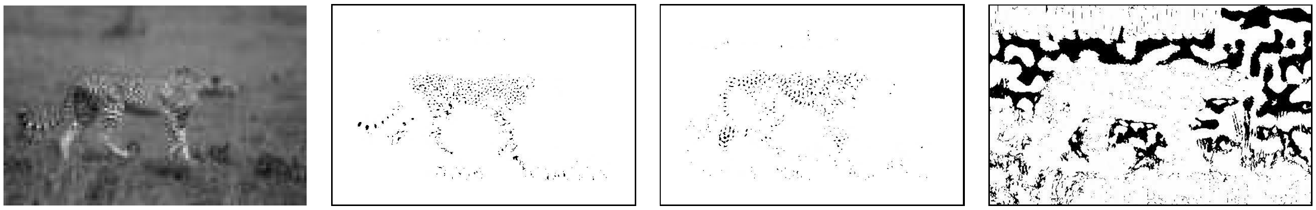

Figure 1. "Generic" descriptors for the cheetah image (on the left). The two images in the middle display "generic" descriptors which characterize the cheetah. The image on the right represents a "generic" descriptor of the background. A "generic" descriptor is represented by the image locations at which it is most probable for the greyvalue descriptor.

$$\mathbf{v}_l = \begin{pmatrix} P(C_1|\mathbf{w}_l) \\ P(C_2|\mathbf{w}_l) \\ \dots \\ P(C_k|\mathbf{w}_l) \end{pmatrix} = \begin{pmatrix} \frac{|\{C^*(\mathbf{p})=C_1|\mathbf{p}\in\mathbf{w}_l\}|}{|\mathbf{w}_l|} \\ \frac{|\{C^*(\mathbf{p})=C_2|\mathbf{p}\in\mathbf{w}_l\}|}{|\mathbf{w}_l|} \\ \dots \\ \frac{|\{C^*(\mathbf{p})=C_k|\mathbf{p}\in\mathbf{w}_l\}|}{|\mathbf{w}_l|} \end{pmatrix}$$

where $\mathbf{w}_l$ is the window centered on the pixel location $\mathbf{p}_l$ and $|\mathbf{w}_l|$ the number of pixels in the window. Note that the "generic" descriptor of the center location is not included, as it is used to compute the conditional joint probability.

We then form sets of frequency vectors $\mathbf{v}_l$ with the same center label (same most probable cluster at the center). Each set represents the conditional joint probability of frequencies with respect to the center $P(\mathbf{v}_l|C^*(\mathbf{p}_l) = C_i)$. The corresponding set of frequency vectors is denoted by V_i. The distribution of V_i is multi-modal and the different modes of the distribution are described by a set of clusters $\{V_{ij}\}$. We use the k-means algorithm to obtain these clusters (cf. section 2.2). Each cluster represents statistically similar neighborhoods. These clusters are in the following referred to as "spatial-frequency" clusters.

3.2. Probability of a spatial-frequency cluster

We now define the probability of a "spatial-frequency" cluster V_{ij}. For an image location $\mathbf{p}_l$, this probability $P(V_{ij}|\mathbf{p}_l)$ is given by $P(V_{ij}|\mathbf{v}_l \wedge \mathbf{d}_l)$:

$$P(V_{ij}|\mathbf{v}_l \wedge \mathbf{d}_l) = \frac{P(\mathbf{v}_l \wedge \mathbf{d}_l|V_{ij})P(V_{ij})}{P(\mathbf{v}_l \wedge \mathbf{d}_l)}$$

$$= \frac{P(\mathbf{v}_l|\mathbf{d}_l \wedge V_{ij})P(\mathbf{d}_l|V_{ij})P(V_{ij})}{\sum_i \sum_j P(\mathbf{v}_l|\mathbf{d}_l \wedge V_{ij})P(\mathbf{d}_l|V_{ij})P(V_{ij})} \qquad (3)$$

We assume in the following that $P(V_{ij})$ are equal and that the distribution of a "spatial-frequency" cluster is approximated with a Gaussian $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We have $P(\mathbf{d}_l|V_{ij}) = P(\mathbf{d}_l|C_i)$ and $P(\mathbf{v}_l|\mathbf{d}_l \wedge V_{ij})$ is defined by :

$$P(\mathbf{v}_l|\mathbf{d}_l \wedge V_{ij}) =$$

$$\begin{cases} \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_{ij}|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\mathbf{d}_l - \boldsymbol{\mu}_{\mathbf{ij}})^t \boldsymbol{\Sigma}_{ij}^{-1}(\mathbf{d}_l - \boldsymbol{\mu}_{\mathbf{ij}})] & \text{if } C^*(\mathbf{d}_l) = C_i \\ 0 & \text{otherwise} \end{cases}$$

Note that we need to evaluate $P(V_{ij}|\mathbf{v}_l \wedge \mathbf{d}_l)$ only if $C^*(\mathbf{d}_l) = C_i$, otherwise its value is zero.

To compute the significance as well as the retrieval score we select for each image location $\mathbf{p}_l$ the most probable "spatial-frequency" cluster, that is the one with the maximum probability $P(V_{ij}|\mathbf{p}_l)$:

$$V^*(\mathbf{p}_l) = V^*(\mathbf{v}_l \wedge \mathbf{d}_l) = \underset{V_{ij}}{\text{argmax}}\, p(V_{ij}|\mathbf{v}_l \wedge \mathbf{d}_l) \qquad (4)$$

4. Significance

The significance or distinctiveness of each "spatial-frequency" cluster allows to determine its importance for the model. We can categorize "spatial-frequency" clusters as positive and distinctive, positive and not distinctive, background (non relevant parts of the positive sample images) and negative.

We want to identify clusters which are positive and distinctive. Intuitively, "spatial-frequency" clusters which appear often in the positive examples and rarely in the negative samples fall into this category. This is captured by our significance measure defined in the following. Note that it is fundamental to keep the non-significant clusters in the model. These clusters are matched to descriptors of the background or negative images and the significance measure allows to eliminate them without using an arbitrary threshold. This avoids false positive responses for test images which do not contain the model. The importance of negative clusters has been confirmed by Sung and Poggio [17] in the context of learning the distribution of global face patches.

In the following we determine which of the "spatial-frequency" clusters are significant for the model. For each cluster V_{ij} we compute its probability for the positive and negative sample images separately. Given a set of m sample images, for which the probabilities are assumed independent and equal ($P(I_j) = 1/m$), we obtain :

$$P(V_{ij}|\{I_1, I_2...I_m\}) = \frac{1}{m}\sum_{q=1}^{m} P(V_{ij}|I_q) \qquad (5)$$

To compute the probability of a "spatial-frequency" cluster for an image, we assume the n pixel locations $\mathbf{p}_l$ to be independent and equally probable ($P(\mathbf{p}_l) = 1/n$). The pixel locations $\mathbf{p}_l$ are described by the descriptors $\mathbf{d}_l$ and the spatial frequencies $\mathbf{v}_l$ over neighborhoods :

$$P(V_{ij}|I) = P(V_{ij}|\{\mathbf{p}_1, \mathbf{p}_2, \dots \mathbf{p}_n\}) = \frac{1}{n}\sum_{l=1}^{n} P(V_{ij}|\mathbf{p}_l) = \qquad (6)$$

$$\frac{1}{n}\sum_{l=1}^{n} \begin{cases} P(V_{ij}|\mathbf{p}_l) & \text{if } V^*(\mathbf{p}_l) = V_{ij} \\ 0 & \text{otherwise} \end{cases}$$

Note that we only include the probability of the most probable "spatial-frequency" cluster. This avoids the accumulation of insignificant probabilities and corresponds to the retrieval algorithm which takes into account only the most probable cluster. The above equations allow to compute the probability of a cluster V_{ij} for a set of positive sample images $P(V_{ij}|\{I_{pos}\})$ as well as for a set of negative sample images $P(V_{ij}|\{I_{neg}\})$. The significance of cluster V_{ij} for a model M is then defined as follows :

$$\mathbf{Sig}(V_{ij}|M) = \frac{P(V_{ij}|\{I_{pos}\})}{P(V_{ij}|\{I_{pos}\}) + P(V_{ij}|\{I_{neg}\})}$$

The values of this significance measure vary between 0 and 1. If the value is close to one, the "spatial-frequency" cluster is significant, that is relevant for the model. For example a "spatial" cluster which has close to zero probability in the negative images and high probability in all or most of the positive examples is significant.

5. Retrieving images

In the previous sections we have constructed a model M from a set of positive and negative images. This model is described by a set of "generic" descriptors, a set of "spatial-frequency" clusters and the significance of each "spatial-frequency" cluster. In the following we want to retrieve images which contain instances of the model as well as localize instances of the model in the images. This issue has been recently addressed for example in the context of face detection [16, 18].

We retrieve and localize instances of a model using a probabilistic score. The first step is to compute the model probability for an individual pixel $P(M|\mathbf{p}_l)$. This probability uses only the most probable "generic" descriptor and the most probable "spatial-frequency" cluster. $P(M|\mathbf{p}_l)$ is determined as follows :

1. For pixel location $\mathbf{p}_l$ we compute its descriptor $\mathbf{d}_l$.
2. For descriptor $\mathbf{d}_l$ we obtain the probabilities $P(C_i|\mathbf{d}_l)$ using equation (1). We then determine the most probable cluster $C^*(\mathbf{d}_l)$ as described by equation (2).
3. The spatial-frequency descriptor $\mathbf{v}_l$ is computed for the neighborhood of pixel $\mathbf{p}_l$. Note that for each pixel in the neighborhood, the most probable "generic" descriptor has to be determined. It is given by the label image.
4. The probabilities $P(V_{ij}|\mathbf{v}_l \wedge \mathbf{d}_l)$ are computed using equation (3) and the most probable "spatial-frequency" cluster $V^*(\mathbf{v}_l \wedge \mathbf{d}_l)$ is determined as described by equation (4).
5. If $\mathbf{Sig}(V^*(\mathbf{p}_l)|M)$ is below a threshold t, the probability $P(M|\mathbf{p}_l)$ is set to zero. t equals 0.5 in our experiments, that is $\mathbf{p}_l$ is rejected if it is more likely to belong to a negative sample.
6. The score of a pixel is computed by $P(M|\mathbf{p}_l) = P(C^*(\mathbf{p}_l)|\mathbf{d}_l)P(V^*(\mathbf{p}_l)|\mathbf{v}_l \wedge \mathbf{d}_l)\mathbf{Sig}(V^*(\mathbf{p}_l)|M)$

For retrieval we determine the probability of a model given an image. If the n pixel locations $\mathbf{p}_l$ are assumed independent and equivalently probable, this probability can be computed by :

$$P(M|I) = P(M|\{\mathbf{p}_1, \mathbf{p}_2, \dots \mathbf{p}_n\}) = \frac{1}{n}\sum_{l=1}^{n} P(M|\mathbf{p}_l)$$

Note that the above equation summarizes pixel-based probability scores. This assumes independence of pixel locations which is in general not valid. We should obtain different scores if significant pixels are spread out over the image or localized in a region. This should be taken into account when computing our score and is currently under investigation.

To localize instance of models in images, we select pixels with high probabilities (see for example figure 4). Selecting such pixels is only a crude method which can easily be improved, for example by including region segmentation [12]. The results are however already more than satisfactory.

6. Experimental results

For our experimental results we constructed models from 15 sample images (5 positive and 10 negative). This corresponds to a realistic setting where negative examples are more easily available. The number of "generic" descriptors was set to 50 and the number of spatial clusters for one "generic" descriptor was set to 10. The size of the neighborhood window was 21x21.

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. Retrieval results are evaluated by computing precision as a function of recall. Precision is the number of relevant

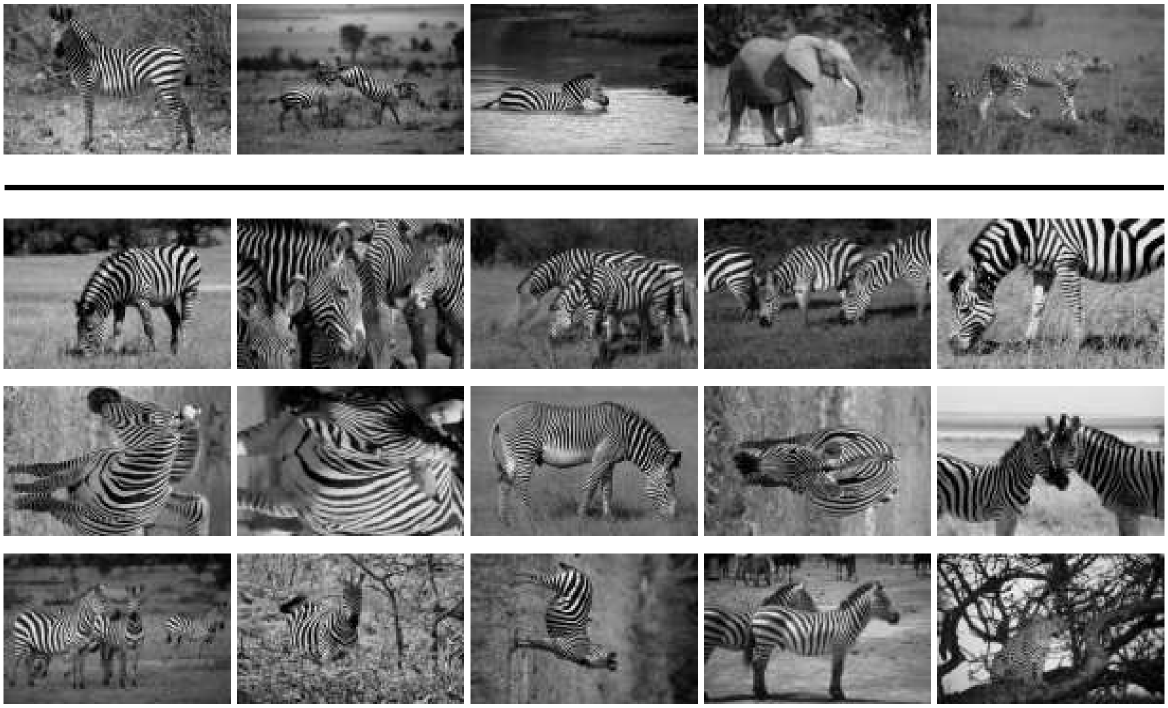

Figure 2. Retrieval results. The top row shows a subset of the training images (3 positive and 2 negative examples). The bottom rows show the first 15 retrieved images ordered by their score (from left to right and from top to bottom).

images retrieved relative to the total number of retrieved images. Recall is the number of relevant images retrieved relative to the total number of relevant images in the database.

The top row of figure 2 shows a subset of the training images (3 positive and 2 negative examples) used to learn the zebra model. The remaining rows display the 15 retrieved images 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. The precision/recall graph is shown in figure 3. Results are comparable to or better than those of other systems which manually extract objects. Moreover, our method allows to localize the model in a retrieved test image by selecting locations with a high score. Results of localizing the zebra model are presented in figure 4. 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.

Results for the cheetah model are displayed in figures 5 and 6. The graph for precision/recall is similar to the one obtained for the zebra model. The three cheetahs are correctly localized in figure 6. Equivalent results were obtained for the giraffe model. They are not shown due to space limitations. Results for faces are displayed in figure 7. The graph for precision/recall is equivalent to those for the "tex-

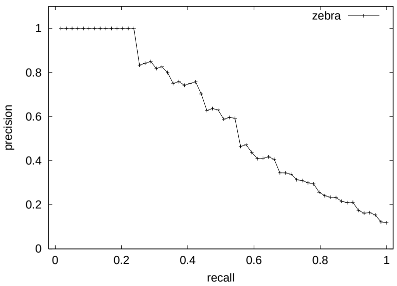

Figure 3. Precision as a function of recall for the zebra model.

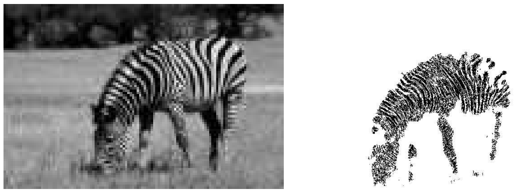

Figure 4. Localization of the zebra model for one of the test images (left). Locations with the high probability scores are displayed in black (right).

tured” animals. Our method is therefore appropriated for textured objects as well as for highly structured ones.

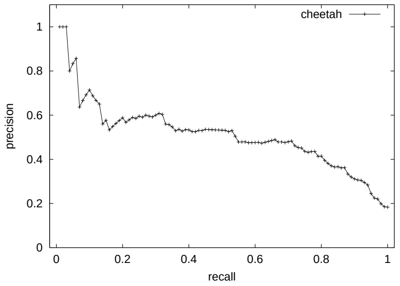

Figure 5. Precision as a function of recall for the cheetah model.

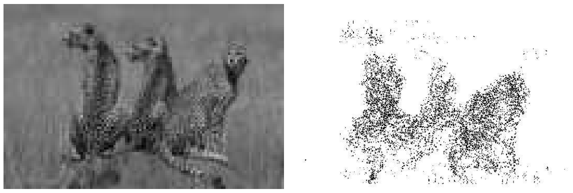

Figure 6. Localization of the cheetah model for one of the test images (left). Locations with high probability scores are displayed in black (right).

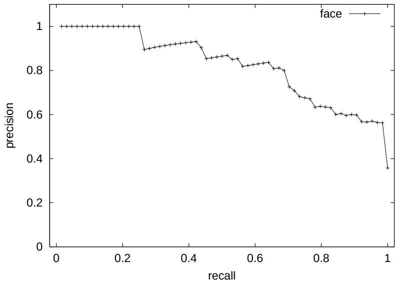

Figure 7. Precision as a function of recall for the face model.

7. Conclusion and discussion

We have presented a novel approach for model construction which significantly improves on the state of the art. It presents the following three advantages. The first is our model representation which captures efficiently “texture-like” visual structure. The second is the learning algorithm which is unsupervised and therefore does not require manual extraction of objects or features. Furthermore, it allows to learn an appropriate representation of the model. The third is the independence of region segmentation and feature extraction which are never perfect.

Finally, we mention four extensions which we are currently investigating. The first is to learn which components of our multi-valued descriptors are significant, that is most appropriate to describe the object. The second is to improve the clustering algorithm and to automatically select the number of clusters. The third is to include global constraints, for example by modeling relations between parts [4] or by segmenting regions [12]. The fourth extension is to add relevance feedback, that is to improve the model over time by user interaction.

References

[1] Y. Amit and D. Geman. A computational model for visual selection. *Neural Computation*, 11(7):1691–1715, 1999.

[2] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In *ICCV*, pp. 675–682, 1998.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[4] D.A. Forsyth and M.M. Fleck. Body plans. In *CVPR*, pp. 678–683, 1997.

[5] D. Gabor. Theory of communication. *Journal I.E.E.*, 3(93):429 – 457, 1946.

[6] A.K. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[7] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 55:367–375, 1987.

[8] S. Konishi and A.L. Yuille. Statistical cues for domain specific image segmentation with performance analysis. In *CVPR*, pp. 125–132, 2000.

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

[10] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions : Cue integration in image segmentation. In *ICCV*, pp. 918–925, 1999.

[11] 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.

[12] N. Paragios and R. Deriche. Geodesic active regions for supervised texture segmentation. In *ICCV*, pp. 926–932, 1999.

[13] A.L. Ratan, O. Maron, W.E.L. Grimson, and T. Lozano-Prez. A framework for learning query concepts in image classification. In *CVPR*, pp. 423–429, 1999.

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

[15] Y. Rubner and C. Tomasi. Texture-based image retrieval without segmentation. In *ICCV*, vol. 2, pp. 1018–1024, 1999.

[16] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *CVPR*, vol. 1, pp. 746–751, 2000.

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

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