A Structured Probabilistic Model for Recognition

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

Abstract

In this paper we derive a probabilistic model for recognition based on local descriptors and spatial relations between these descriptors. Our model takes into account the variability of local descriptors, their saliency as well as the probability of spatial configurations. It is structured to clearly separate the probability of point-wise correspondences from the spatial coherence of sets of correspondences. For each descriptor of the query image, several correspondences in the image database exist. Each of these point-wise correspondences is weighted by its variability and its saliency. We then search for sets of correspondences which reinforce each other, that is which are spatially coherent. The recognized model is the one which obtains the highest evidence from these sets.

To validate our probabilistic model, it is compared to an existing method for image retrieval. The experimental results are given for a database containing more than 1000 images. They clearly show the significant gain obtained by adding the probabilistic model.

1 Introduction

Recognition is a recurrent topic in computer vision. Existing methods can be divided into two categories. The first one uses global descriptors, that is descriptors are computed for the entire image [10, 18]. The second one uses local characteristics and spatial relations between them [3, 6, 11, 13].

The second category is more appropriate to recognize a query image which can be a part of a model image or can be modified (rotated, scaled, seen from a different viewpoint, ...). Furthermore, methods of this category are robust to occlusion and varying background and allow recognition of multiple objects.

In this paper we present a probabilistic model for the second category. It is generic, that is completely independent of the local characteristics as well as the nature of the spatial relations. It takes into account the variability of local descriptors, their saliency as well as the probability of spatial configurations. It is structured to clearly separate the probability of point-wise matches, the density of the descriptors in the database and the spatial coherence of sets of matches.

Using a probabilistic model has several advantages. Firstly, it makes recognition more robust as uncertainties are naturally integrated into the model. Secondly, it avoids immediate decisions that is the model is independent of intermediate thresholds. This allows to make irreversible commitments as late as possible. Moreover, different weights can be attributed to individual characteristics.

Recently probabilistic methods have been used to determine best candidates. However, none of the existing methods completely models the recognition process (cf. section 1.1). Moreover, they don't show the gain obtained by adding a probabilistic model. This gain is measured here; experiments compare the addition of a probabilistic model to an existing method [13]. This method computes local photometric signatures (greyvalue invariants) at interest points. Spatial coherence is based on comparing angles of neighboring points.

1.1 Previous Work

Several authors have developed probabilistic models for methods based on global descriptors. For example, Moghaddam [8] estimates the density of his eigenspace decomposition to formulate a maximum likelihood estimation. This density is estimated by a multivariate mixture of Gaussian models. Sung [17] is another example for such an approach. He learns a representation for faces by estimating the distribution of the pixel values in face regions. The use of only one descriptor allows important simplifications which make such methods not applicable in the case of local descriptors.

Schiele as well as Mohr [12, 9] use the frequency of their local descriptors to determine the most probable model. However, they neither integrate the distance of correspondences nor use any spatial coherence.

Burl [2] as well as Schneiderman [15] compute the saliency of descriptors for a given model. They can recognize a generalized model learned from a big set of test images. They randomly search in the query

image all occurrences of their descriptors. Such methods are not scalable to recognition of multiple models. Moreover, no spatial relations are used. Similarly, MacCormick [7] computes descriptors along object boundaries and adds a spatial relation between descriptors.

Shimshoni [16] uses local geometric descriptors (angles and length ratios). He models the theoretical probability of a correspondence based on the viewing sphere. Furthermore, he computes the pose for a set of correspondences. The probability of a hypothesized model is increased if spatial coherence (pose) is verified. Compared to our model Shimshoni neither models the frequency of his descriptors nor uses the coherence of hypotheses. Moreover, his spatial relation is global and no photometric information is used.

1.2 Overview of the paper

Section 2 describes the probabilistic model. The existing retrieval algorithm as well as implementation details for the probabilistic version are presented in section 3. Section 4 shows the gain obtained when comparing the probabilistic model to the existing retrieval algorithm. In section 5 we discuss the potential extensions of this work.

2 Probabilistic model

2.1 Definition and notations

A query image is in the following denoted by Q and the set of model images by $\mathcal{M} = \{M_m\}$ with $\hat{m}$ being the number of model images. Each image I (either a query or a model image) is characterized by a set of descriptors $\mathcal{D}(I) = \{D_d(I)\}$ with $\hat{d}$ being the number of descriptors for the image I. Each descriptor $D_i(I)$ is computed at a corresponding image location $L_i(I)$.

For each descriptor of a query image $D_i(Q)$ there exist correspondences (matches) with descriptors in different model images. This set of correspondences is denoted by $\mathcal{C}(D_i(Q)) = \{C_c(D_i(Q))\}$ where C_c is the c-th correspondence of the descriptor $D_i(Q)$. The order in the set is arbitrary. For example $C_1(D_1(Q)) = D_4(M_3)$ is the first element in the ordered set $\mathcal{C}(D_1(Q))$. It indicates that there is a correspondence between the descriptors $D_1(Q)$ and $D_4(M_3)$.

In general, there is more than one correspondence per descriptor. Different matching hypotheses are therefore obtained. A hypothesis is the combination of one correspondence per descriptor of the query image. With $\hat{d}$ being the number of elements in $\mathcal{D}$, a hypothesis is defined by

$$H_h = \{C_{h_1}(D_1(Q)) \wedge \ldots \wedge C_{h_{\hat{d}}}(D_{\hat{d}}(Q))\}$$

where the set $\{h_i\}$ defines the correspondences of a hypothesis H_h. The h_i-th correspondence of each descriptor $D_i(Q)$ is part of the hypothesis.

There are $|\mathcal{C}(D_1(Q))|.|\mathcal{C}(D_2(Q))|. \ \ldots \ .|\mathcal{C}(D_{\hat{d}}(Q))|$ hypotheses. Descriptors with no correspondences are not taken into account. The set of all possible hypotheses $\mathcal{H} = \{H_1 \vee \ldots \vee H_{\hat{h}}\}$ represents a query image where $\hat{h}$ is the number of all hypotheses.

2.2 Probability of a model

Given a query image, we want to find the most similar model image, i.e. the model image with the highest probability $P(M_m|Q)$. If a query image is represented by the set of all possible hypotheses, the probability $P(M_m|Q)$ can be expressed by $P(M_m|\mathcal{H})$.

When using the set of all hypotheses $\mathcal{H}$, the most probable model emerges by the number of evidences for this model. Another possibility is to use the hypothesis with the highest probability. However, experimental results show that the performance of a matching method is not as good as using a set of hypotheses.

Using Bayes theorem, $P(M_m|\mathcal{H})$ can be rewritten by :

$$P(M_m|\mathcal{H}) = \frac{P(\mathcal{H}|M_m)\,P(M_m)}{P(\mathcal{H})}$$

We can assume that all hypotheses and all models are independent and then obtain :

$$P(M_m|\mathcal{H}) = \frac{\sum_{h=1}^{\hat{h}} P(H_h|M_m)\,P(M_m)}{\sum_{h=1}^{\hat{h}} P(H_h)}$$

$$\text{where } \; P(H_h) = \sum_{m=1}^{\hat{m}} P(H_h|M_m)\,P(M_m)$$

The probability of all model images $P(M_m)$ can be assumed equal. The key problem is therefore to compute $P(H_h|M_m)$. This probability should take into account :

- the global coherence of the hypothesis ($P_{\mathbf{h}}$)
- the quality of its correspondences ($P_{\mathbf{c}}$)
- the spatial coherence of the corresponding descriptors ($P_{\mathbf{s}}$)

The probability $P(H_h|M_m)$ is then the product of three (independent) probabilities :

$$P(H_h|M_m) = P_{\mathbf{h}}(H_h|M_m).P_{\mathbf{c}}(H_h|M_m).P_{\mathbf{s}}(H_h|M_m)$$

The computation of these three probabilities are detailed in the following sections.

2.3 Probability of a hypothesis

The probability of a hypothesis measures the global coherence of a hypothesis. A hypothesis is more coherent if it involves fewer models. If it involves many models, it is likely that this hypothesis contains correspondences due to clutter or false correspondences. This probability is independent of a given model.

$$P_{\mathbf{h}}(H_h|M_m) = P_{\mathbf{h}}(H_h) = e^{-\frac{\#models}{\beta_h}}$$

where $\#models$ is the number of different models in H_h and β_h is a normalization factor

2.4 Probability of correspondences

The probability of a correspondence depends on the similarity of the matched descriptors as well as on their saliency. The similarity measure is based on the distance of the two corresponding descriptors. Saliency expresses the frequency of the descriptors. The smaller the frequency, the higher is the probability of a correspondence.

The different correspondences of a hypothesis are independent. We then obtain for $P_{\mathbf{c}}(H_h|M_m)$:

$$P_{\mathbf{c}}(H_h|M_m) = \prod_{i=1}^{\hat{d}} P_{\mathbf{c}}\Big(C_{h_i}\big(D_i(Q)\big)|M_m\Big)$$

The probability of a correspondence given a model $P_{\mathbf{c}}\Big(C_{h_i}\big(D_i(Q)\big)|M_m\Big)$ can only be computed if the correspondence matches a descriptor of M_m. The probability of a correspondence is then the product of two (independent) probabilities (similarity $P_{\mathbf{c_s}}$ and frequency $P_{\mathbf{c_f}}$). If the correspondence represents a false match, it is assigned the probability of a false match ϵ_c. We then express $P_{\mathbf{c}}(C_k(D_i(Q))|M_m)$ by:

$$P_{\mathbf{c}}(C_k(D_i(Q))|M_m) = \begin{cases} P_{\mathbf{c_s}}\Big(D_i(Q), D_j(M_m)\Big) \cdot P_{\mathbf{c_f}}(D_i(Q)|M_m) & \text{if } C_k(D_i(Q)) \in \mathcal{D}(M_m) \\ \epsilon_c & \text{otherwise} \end{cases}$$

The probability of similarity should take into account the statistical distribution of a descriptor. We have chosen to use the Mahalanobis distance $dist_M$. For two descriptors D_i and D_j, this distance is defined by $dist_M(D_i, D_j) = \sqrt{(D_i - D_j)^T \Lambda^{-1} (D_i - D_j)}$. The covariance matrix Λ takes into account the variability of the descriptors, i.e. their uncertainty. The covariance matrix is estimated from the data as described in section 3.1.

The probability of similarity is then defined by:

$$P_{\mathbf{c_s}}(D_i(Q), D_j(M_m)) = e^{-\frac{dist_M(D_i(Q), D_j(M_m))}{\beta_c}}$$

where β_c is a normalization factor

The probability of frequency for a descriptor is the ratio of the number of the correspondences with the model M_m over the total number of correspondences. It is defined by:

$$P_{\mathbf{c_f}}(D_i(Q)|M_m) = \frac{|\mathcal{C}(D_i(Q)) \cap \mathcal{D}(M_m)|}{|\mathcal{C}(D_i(Q))|}$$

2.5 Probability of configurations

The probability of configurations computes the similarity of spatial configurations between a hypothesis and a model. This probability can be determined from global or semi-local configurations. To be robust to projective transformations, semi-local configurations have been used, since they are a first order approximation of a projective transformation.

Probabilities of different image locations are not independent, but are assumed to be for simplicity of computation. We then obtain

$$P_{\mathbf{s}}(H_h|M_m) = \prod_{i=1}^{\hat{d}} P_{\mathbf{s}}\Big(C_{h_i}\big(D_i(Q)\big)|M_m\Big)$$

Given a correspondence $C_k(D_i(Q))$ and a model M_m, the spatial configuration is defined by the n neighbors in the subset of correspondences with the model M_m. The probability is defined by the similarity of the angular structure of these n neighbors between the query image and the model image. The image locations $L_i(Q)$ and $L_i(M_m)$ are used to compute the angles. The probability $P_{\mathbf{s}}(C_k(D_i(Q))|M_m)$ is defined by

$$P_{\mathbf{s}}(C_k(D_i(Q))|M_m) = \begin{cases} \prod_{l=1}^{n} P\Big(\alpha_l(D_i(Q)), \alpha_l(D_j(M_m))\Big) & \text{if } C_k(D_i(Q)) \in \mathcal{D}(M_m) \\ \epsilon_s & \text{otherwise} \end{cases}$$

where ϵ_s is the spatial probability of a false correspondence

The probability that two angles are the same is

$$P(\alpha_l(D_i(Q)), \alpha_l(D_j(M_m)) = e^{-\frac{|\alpha_l(D_i(Q)) - \alpha_l(D_j(M_m))|}{n\,\beta_s}}$$

n is used for normalization such that $P_{\mathbf{s}}$ and $P_{\mathbf{c}}$ are of the same magnitude. β_s is a normalization factor.

3 Image retrieval algorithms

In this section we first present an existing method for recognition based on a voting algorithm. This method is briefly described (for more details see [13]). We then explain how to replace the voting mechanism by our probabilistic model. This induces a small change on the use of spatial configurations. Implementation details are given.

3.1 Voting approach

Local intensity invariants are used as image descriptors. These invariants are computed at interest points. Such an approach allows to differentiate between many objects in the case of partial visibility, similarity transformations, extraneous features, and small perspective deformations. The steps of the algorithm are detailed in the following:

1) Image locations are interest points which are automatically extracted using the Harris detector [4]. The basic idea of this detector is to use the autocorrelation function in order to determine locations where the signal changes in two directions. A comparison of different detectors under varying conditions [14] has shown best results for the Harris detector.

2) Descriptors are 9 dimensional vectors of intensity invariants which describe the local neighborhood of an interest point. Intensity invariants are combinations of local intensity derivatives [5]. The invariants used here are limited to third order. To deal with scale changes, the vector of invariants is computed at several scales [19]. Our characterization is now invariant to similarity transformations which are additionally quasi-invariant to 3D projections [1].

3) Correspondences are determined by comparing descriptors. Similarity of two invariant vectors is quantified using the Mahalanobis distance $dist_M$. Descriptors with a distance measure below a threshold correspond. In order to obtain accurate results for the distance, it is important to have a representative covariance matrix Λ which takes into account signal noise, luminance variations, as well as imprecision of the interest point location. As a theoretical computation seems impossible to derive given realistic hypotheses, it is estimated statistically here by tracking interest points in image sequences.

To avoid an exhaustive comparison of the descriptors of the query image with all the descriptors of the database, an indexing technique is used. This technique is based on a variant of a k-d tree. It leads to a very efficient recognition. The mean retrieval time for our database containing 1020 objects (see figure 1) is less than 5 seconds on a UltraSparc 30.

4) Spatial coherence is used to reduce the number of correspondences. A correspondence is only kept if a percentage of neighboring points matches and respects the geometric constraints (angles). For our experiments, the number of neighbouring points n is set to 5 and angles are geometrically consistent if they don't vary by more than 0,2 radian. We require that at least 50% of the neighbors respect these constraints.

5) Voting selects the most similar model for a given query image Q. The idea is to sum over the number of times each model is selected in the set of correspondences. For each model M_k, $T(k)$ represents this sum:

$$T(k) = \sum_{i=1}^{\hat{d}} \sum_{j=1}^{|\mathcal{C}(D_i(Q))|} \begin{cases} 1 \text{ if } C_j(D_i(Q)) \in \mathcal{D}(M_k) \\ 0 \text{ otherwise} \end{cases}$$

$\hat{d}$ is the number of descriptors of the query image and $|\mathcal{C}(D_i(Q))|$ is the number of correspondences for a given descriptor.

To obtain a comparable measure the vector $T(k)$ is normalized by the total number of votes ($\sum_{k=1}^{\hat{m}} T(k)$). The model that is selected most often, that is the model with the highest score $T(k)$, is considered to be the most similar model image.

3.2 Probabilistic approach

The probabilistic approach uses the same image locations (1), descriptors (2) and correspondences (3) as the approach described in the previous section. Spatial coherence (4) are no longer used to reduce the number of correspondences, but are integrated in the computation of probabilities. The probabilities are computed as described in section 2. The most similar model is the one with the highest probability.

The normalization factors necessary for the computation of the probabilities have been estimated over a large set of test images. β_h is estimated to 30 and β_s to 8 degrees. β_c equals the threshold of the Mahalanobis distance which is set to 6. The ϵ_c and ϵ_s are set to $1/e$ which corresponds to $dist_M(D_i(Q), D_j(M_m)) = \beta_c$ and $|\alpha_l(D_i(Q)) - \alpha_l(D_j(M_m))| = \beta_s$.

Due to the large number of hypotheses it is not feasible to compute the probabilities of all hypotheses. On average 100 points are extracted for a query image and 10 correspondences are obtained for each point, that is there are 10^{100} hypotheses. We therefore use a heuristic to select hypotheses. For our experiments 2000 hypotheses are examined. Experiments have shown that this number is sufficient. Using more hypotheses does not significantly influence the results, that is scores and ranking order of the models. Yet, selection of the hypotheses has shown to be very important. Hypotheses which aren't selected contain mainly wrong correspondences and therefore don't increase the probability of any model.

4 Experimental results

This section displays the gain obtained by adding a probabilistic model. Experimental results for a large database clearly show the improvement: the recognition rate increases for a set of test images and the scores are more distinctive.

4.1 Experimental setup

The database used for our experiments contains 1020 images : 200 paintings, 100 aerial images and 720 images of 3D objects (see figure 1). 3D objects include the Columbia database. In the case of a planar object, an object is represented by one image in the database. This is also the case for nearly planar objects as for aerial images. A 3D object has to be represented by images taken from different viewpoints. Images are stored in the database with 20 degrees viewpoint changes.

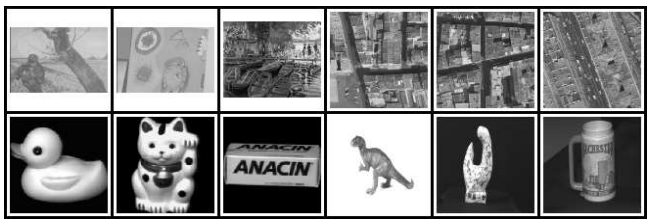

Figure 1: Some images of the database. The database contains 1020 images.

The test set consists of 100 aerial images taken from a different point of view. The left image of figure 2 shows an image of the test set and the right image is the corresponding image of the database. Notice that buildings appear differently because viewing angles have changed and cars have moved. Another test

Figure 2: The left image shows an image of the test set ; the right one is the corresponding image of the database. Notice that buildings appear differently because viewing angles have changed and cars have moved (courtesy of Istar).

image is shown in figure 3. A 3D object is in front of cluttered background and the pose is different from the one stored in the database (by 10 degrees).

4.2 Comparison of the two methods

Experiments first show that the recognition rate increases for the test set of 100 aerial images (taken from

Figure 3: A 3D object in front of cluttered background.

a different viewpoint). Parts of different sizes are extracted from these images ; the size varies between 100 and 10 percent of the surface of the image. The image number and the part position of the query images are determined randomly. For each size 100 query images are used. Parts with less than 5 interest points are excluded.

Figure 4 compares the two methods : the existing method ("vote") and the probabilistic approach ("proba"). The recognition rate is given for different sizes of the query image. This recognition rate is the percentage of correct retrieval over the total number of query images (100). Correct retrieval means that the most similar model is the one which corresponds to the query image.

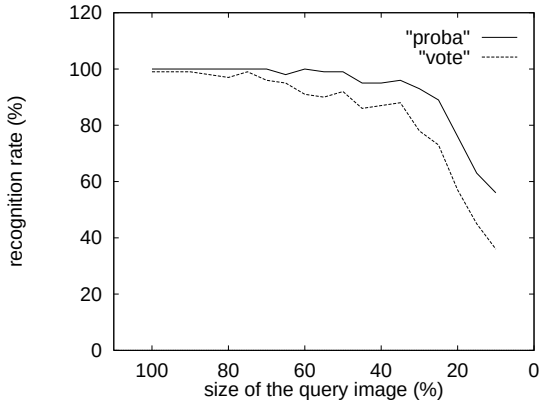

Figure 4: Comparison of the two methods : the existing method ("vote") and the probabilistic approach ("proba"). The recognition rate is displayed as a function of the size of the query image. Results of the probabilistic method are always better than those of the existing approach. The gain is more significant for small parts.

Figure 4 clearly shows the improvement obtained when using the probabilistic method. For small parts

the gain is more significant. For these parts there are fewer points and the weighting of the points is then crucial.

Figure 5 shows that the scores obtained by the probabilistic method are more distinctive. The upper table displays the results obtained for one of the aerial test images (the left image of figure 2) ; the lower table shows the results for the image of figure 3. For both tables the left column ("vote") gives the results for the existing approach and the right one ("proba") for the probabilistic approach. The image numbers and its corresponding scores are displayed in decreasing order. Only the results of the four best scores are shown. Both methods correctly recognize image 24 and 948/949 respectively. 948 and 949 are the two closest views (+/- 10 degrees) for the dinosaur image. The score of the probabilistic method is clearly more distinctive. This makes such a score more appropriate for automatic thresholding. This is very important if we don't know whether the object is contained in the database and in the presence of multiple objects in the scene.

vote		proba	
image number	score	image number	score
24	0.176	24	0.724
143	0.024	182	0.011
0	0.019	0	0.010
75	0.017	20	0.010

vote		proba	
image number	score	image number	score
948	0.103	949	0.716
144	0.049	948	0.155
949	0.044	965	0.072
143	0.039	964	0.030

Figure 5: Comparing the scores of the two methods : the existing method ("vote") and the probabilistic method ("proba"). The query image for the upper table is the left image of figure 2 and for the lower table the image of figure 3. The four best scores are shown. Results obtained by the probabilistic method are clearly more distinctive.

5 Conclusion and discussion

We have derived and implemented a structured probabilistic model for recognition. This model integrates the uncertainty of the data as well as its distinctiveness. Results improve significantly when adding our probabilistic model to an existing recognition algorithm.

The presented model seems appropriate to recognize classes of objects. In this case it is important to capture the appropriate variability of a descriptor and to use the saliency of descriptors to select the ones which best represent an object class. The probabilistic model can also be used to select the most salient image locations. Other potential extensions include a reduction of the size of the database by using only the most discriminant points.

References

[1] T.O. Binford and T.S. Levitt. Quasi-invariants: Theory and exploitation. In DARPA *Image Understanding Workshop*, 1993.

[2] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, 1998.

[3] D.A. Forsyth and M.M. Fleck. Body plans. In *CVPR*, 1997.

[4] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

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

[6] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C.v.d. Malsburg, R.P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *Transactions on Computers*, 42(3):300–311, 1993.

[7] J. MacCormick and A. Blake. Spatial dependence in the observation of visual contours. In *ECCV*, 1998.

[8] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. In *ICCV*, 1995.

[9] R. Mohr, S. Picard, and C. Schmid. Bayesian decision versus voting for image retrieval. In *CAIP*, 1997.

[10] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *IJCV*, 14:5–24, 1995.

[11] R.P.N. Rao and D.H. Ballard. Object indexing using an iconic sparse distributed memory. In *ICCV*, 1995.

[12] B. Schiele and J..L. Crowley. Probabilistic object recognition using multidimensional receptive field histogram. In *ICPR*, 1996.

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

[14] C. Schmid, R. Mohr, and Ch. Bauckhage. Comparing and evaluating interest points. In *ICCV*, 1998.

[15] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *CVPR*, 1998.

[16] I. Shimshoni and J. Ponce. Probabilistic 3D object recognition. In *ICCV*, 1995.

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

[18] M.J. Swain and D.H. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991.

[19] A.P. Witkin. Scale-space filtering. In *IJCAI*, 1983.