Selection of Scale-Invariant Parts for Object Class Recognition

Gy. Dorkó and C. Schmid
INRIA Rhône-Alpes, GRAVIR-CNRS
655, av. de l'Europe, 38330 Montbonnot, France
{dorko,schmid}@inrialpes.fr

Abstract

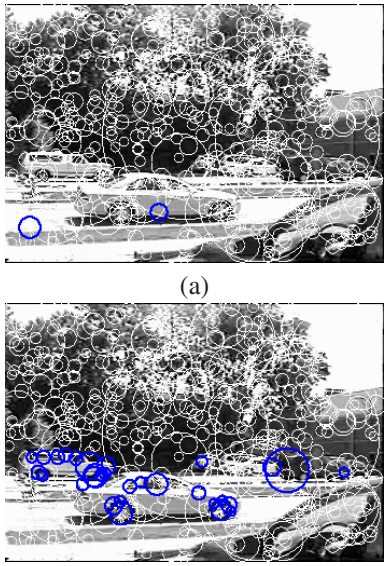
This paper introduces a novel method for constructing and selecting scale-invariant object parts. Scale-invariant local descriptors are first grouped into basic parts. A classifier is then learned for each of these parts, and feature selection is used to determine the most discriminative ones. This approach allows robust part detection, and it is invariant under scale changes—that is, neither the training images nor the test images have to be normalized.

The proposed method is evaluated in car detection tasks with significant variations in viewing conditions, and promising results are demonstrated. Different local regions, classifiers and feature selection methods are quantitatively compared. Our evaluation shows that local invariant descriptors are an appropriate representation for object classes such as cars, and it underlines the importance of feature selection.

1. Introduction

Recognizing general object classes and parts is one of the most challenging problems in computer vision. The combination of computer vision and machine learning techniques has recently led to significant progress [1, 17, 18], but existing approaches are based on fixed-size windows and do not make use of recent advances in scale-invariant local feature extraction [6, 8]. Thus, they require normalizing the training and test images.

We propose in this paper a method for selecting discriminative scale-invariant object parts. Figure 1(a) demonstrates the importance of feature selection. It shows the output of a scale-invariant operator for finding significant circular patches in images [6]. In this context, it is natural to define object parts in terms of clusters of patches with similar brightness patterns. However, consider the two patches marked in black in the figure. The corresponding patterns are very close, but one of the patches lies on a car, while the other lies in the background. This shows that the corresponding part is not discriminative for cars (in this environment at least). To demonstrate the effect of the proposed feature selection method, Fig. 1(b) shows the initially detected features (white) and discriminative descriptors determined by feature selection (black). These are the ones which should be used in a final, robust detection system.



(a) (b)

Figure 1. Illustration of feature selection (see text for details).

1.1. Related Work

Most appearance-based approaches to object class recognition characterize the object by its whole image [9, 15]. They are not robust to occlusion and also suffer

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

from a lack of invariance. Furthermore, these methods are only applicable to rigid objects and either they require preliminary segmentation, or windows have to be extracted for different locations, scales and rotations. The representation is also high-dimensional, therefore many learning techniques cannot be used. To overcome these problems the use of local features is becoming increasingly popular for object detection and recognition.

Weber et al. [18] use localized image patches and explicitly compute their joint spatial probability distribution. This approach has recently been extended to include scale-invariant image regions [11]. Agarwal and Roth [1] first learn a vocabulary of parts, determine spatial relations on these parts and use them to train a Sparse Network of Winnows (SNoW) learning architecture. Lazebnik et al. [5] take advantage of local affine invariants to represent textures.

Some recent methods combine feature selection and local features. Viola and Jones [17] select rectangular (Haar-like) features with an AdaBoost trained classifier. Chen et al. [3] also use this boosting approach for components learned by local non-negative matrix factorization. Amit and Geman [2] employ small, localized and oriented edges and combine them with decision trees. Mahamud and Hebert [7] select discriminative object parts and develop an optimal distance measure for nearest neighbor search. Rikert et al. [12] use a mixture model, but only keep the discriminative clusters and Schmid [14] selects significant texture descriptors in a weakly supervised framework. Both approaches select features based on their likelihood. Ullmann et al. [16] use image fragments and combine them with a linear discriminative type classification rule. Their selection algorithm is based on mutual information.

1.2. Overview

The first step of our approach is the detection of scale-invariant interest points (regions) and the computation of a rotation-invariant descriptor for each region (cf. section 2.1). These descriptors are then clustered and we obtain a set of parts each of which is described by a classifier (cf. section 2.2). Finally, we select a set of discriminative parts/classifiers (cf. section 3). An experimental evaluation compares different approaches to region extraction, classification and selection (cf. Section 4). Finally in Section 5 we conclude and outline our future work.

2. Object-Part Classifiers

In the following we first describe how to compute invariant descriptors and then explain how to learn object part classifiers.

2.1. Scale-Invariant Descriptors

To obtain invariant descriptors we detect scale-invariant interest points (regions) and characterize each of them by a scale, rotation and illumination invariant descriptor.

Scale-invariant detectors. We have used two different scale-invariant detectors: Harris-Laplace [8] and DoG (Difference-of-Gaussian) [6]. Harris-Laplace detects multi-scale Harris points and then selects characteristic points in scale-space with the Laplace operator. DoG interest points [6] are local scale-space maxima of the Difference-of-Gaussian.

The image locations (regions) selected by the two detectors differ: The DoG detector finds blob-like structures and Harris-Laplace detects corners and highly textured points. Examples for detection are shown in the first column of Figure 7.

Scale and rotation invariant descriptors. The output of the two detectors are scale-invariant regions of different sizes. These regions are first mapped to circular regions of a fixed-sized radius. Point neighborhoods which are larger than the normalized region, are smoothed before the size normalization. Rotation-invariance is obtained by rotation in the direction of the average gradient orientation (within a small point neighborhood). Affine illumination changes of the pixel intensities $(aI(\mathbf{x}) + b)$ are eliminated by normalization of the image region with the mean and the standard deviation of the intensities within the point neighborhood. These normalized regions are then described by the SIFT descriptor (Scale Invariant Feature Transform) [6]. SIFT is computed for 8 orientation planes and each gradient image is sampled over a 4x4 grid of locations. The resulting descriptor is of dimension 128.

2.2. Classifiers

Object-part classifiers are learned from sets of similar descriptors. These sets are obtained automatically by clustering local invariant descriptors. Figure 2 shows a few images of two different clusters. The top row displays a "tire" cluster and the bottom row a "front window" cluster.

We have used two types of classifiers: Support Vector Machines (SVMs) and classification based on a Gaussian mixture model (GMM). The training data consists of positive and negative descriptors. Note that the descriptors are labeled manually.

Support Vector Machine. Each object part is described by a separate SVM. A descriptor is classified as a part, if the the corresponding SVM has a positive response.

The SVMs are trained as follows. The first step is to determine groups of similar descriptors. We cluster the positive training descriptors with a hierarchical clustering algorithm. The number of clusters is set to 300. We then

Figure 2. A few images of two different clusters. The first row shows a cluster which represents "tires". The second row shows a cluster which contains regions detected in the "front window".

learn a linear SVM [4] for each positive cluster. The SVM is trained with all descriptors of the positive cluster and a subset of the negative descriptors. This subset are the medians of negative clusters. Note that this pre-selection of the negative samples is necessary. Otherwise we would have highly unbalanced training sets, which can not be handled by current state of the art SVM techniques.

Gaussian mixture model. The distribution of the training descriptors is described by a Gaussian mixture model $\sum_i p(\mathbf{x}|C_i)\, P(C_i)$. Each Gaussian C_i corresponds to an "object-part". A descriptor is assigned to the most likely Gaussian C_i, i.e. it is classified as the corresponding part.

Each $p(\mathbf{x}|C_i)$ is assumed to be a Gaussian with mean μ_i and covariance matrix Σ_i. We use the EM algorithm to estimate the parameters of the mixture model, namely the means μ_i, covariances Σ_i, and mixing weights $P(C_i)$. EM is initialized with the output of the K-means algorithm. In this work, we use the 600 components to describe the training set which includes positive and negative descriptors. We use all positive and randomly choose the same number of negative descriptors. We limit the number of free parameters in the optimization by using diagonal Gaussians. This restriction also helps prevent the covariance matrices from becoming singular.

3. Feature Selection

Given a set of classifiers, we want to rank them by their distinctiveness. Here, we use two different feature selection techniques: likelihood ratio and mutual information. These techniques assign a score to each classifier depending on its performance on a validation set.

The two feature selection methods are based on the probabilities described in the following. Let C_i be a classifier and O the object to be recognized (detected). $P(C_i = 1|O = 1)$ is the probability that C_i classifies a object O descriptor correctly (i.e the true positives for C_i over the number of positives descriptors). $P(C_i = 1|O = 0)$ is the probability of non-objects descriptors being accepted by classifier C_i.

Likelihood ratio. A classifier C_i is representative of an object class if it is likely to be found in the class, but unlikely to be detected in non-class images. The likelihood ratio of classifier C_i is defined by:

$$L(C_i) = \frac{P(C_i = 1|O = 1)}{P(C_i = 1|O = 0)}$$

Mutual information. Mutual information [10] selects informative features. Mutual information $I(C_i, O)$ between the classifier C_i and the object class O is defined by:

$$I(C_i, O) =$$

$$\sum_{c\in\{0,1\}} \sum_{o\in\{0,1\}} P(C_i = c, O = o) \log \frac{P(C_i = c, O = o)}{P(C_i = c)P(O = o)}$$

For both feature selection methods presented above, the higher the score the more relevant it is. The difference between the two methods is illustrated by Figure 3. The image is one of the test images and regions are detected with the DoG detector.

The top 4 rows show the descriptors classified as object parts by the n best SVM classifiers. We can see that the likelihood selects very specific features which are highly discriminative. For example there is no car feature detected

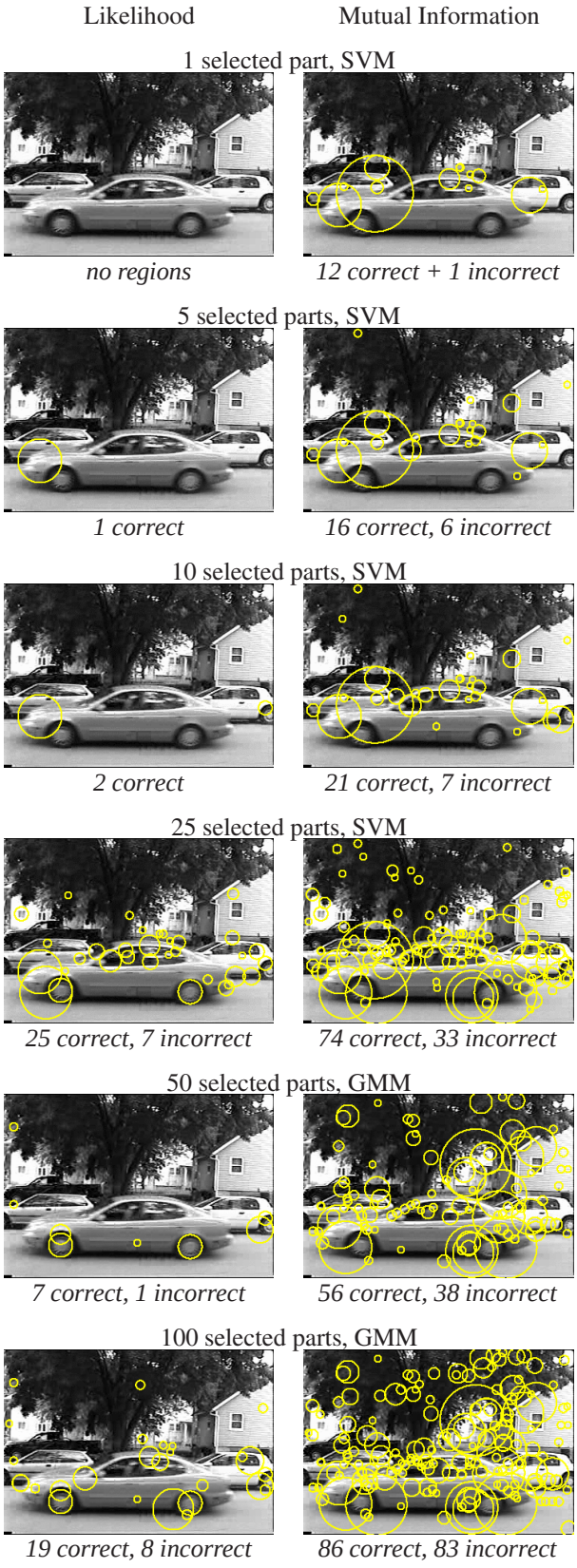


Figure 3. Comparison of feature selection with likelihood ratio and mutual information.

by the "best" classifier in the case of likelihood ratio. This feature is very specific and only detected on certain cars. In contrast mutual information selects informative features. For example the first selected features already classifies 13 descriptors as object descriptors. Note that one of them is incorrect. If we look at the overall performance of the two feature selection methods, we can observe that the likelihood ratio performs slightly better than mutual information. Note however that fewer classifiers are used in the case of mutual information. This is confirmed by the images in the 2 bottom rows which show the results for the n best GMM classifiers as well as by the quantitative evaluation in Section 4. Note that to obtain similar results for GMM we have to use more classifiers. This is due to the fact that there are twice as many classifiers and that they are in general more specific.

4. Experiments

In the previous sections we have presented several techniques for the different steps of our approach. We now evaluate these techniques in the context of car detection. We then present car detection results for a few test images.

4.1. Set-up

Our training database contains 617 images of cars with a relatively large amount of background (more that 50% on average). We have marked the cars in these images manually. Note that the car images can be at different scale levels and do not require normalization. We extract scale-invariant interest points (regions) with the DoG detector and Harris-Laplace. For DoG we obtained 36810 positive and 300998 negative regions. For Harris-Laplace we detected 30631 positive and 161188 negative regions.

The test images were taken independently and contain unseen cars and unseen background. We have used 186 images which often contain several cars and a large amount of background. To evaluate and compare different methods, we marked them manually. We therefore know that the test images contain 15623 positive and 86913 negative descriptor if the DoG detector is used and 13997 and 39890 descriptors for Harris-Laplace.

4.2. Evaluation of Different Methods

In the following we evaluate our approach and compare the performance of different techniques. The comparison criterion is true positive rate (the number of positive descriptors retrieved over the total number of positive descriptors) against false positive rate (the number of false positives over the total number of negatives descriptors).

Classification and Feature selection. Figure 4 compares

the performance of two different classification techniques and two different feature selection criteria. Regions are extracted with the DoG detector. Fig. 4 shows the ROC curve (true positive rate against false positive rate). We can observe that the combination of Gaussian mixture model and likelihood ratio performs best. The second best is the curve for SVM and likelihood ratio which performs slightly better than SVM and mutual information. Results for the combination of mixture model and mutual information are significantly worse. This can be explained by the fact that the classifiers are mostly specific. Fig. 5(a) and (b) compare the criteria true positive rate and false negative rate separately as a function of the number of selected classifiers. As expected mutual information has a higher true positive rate and the false negatives rate is better (lower) for the likelihood ratio.

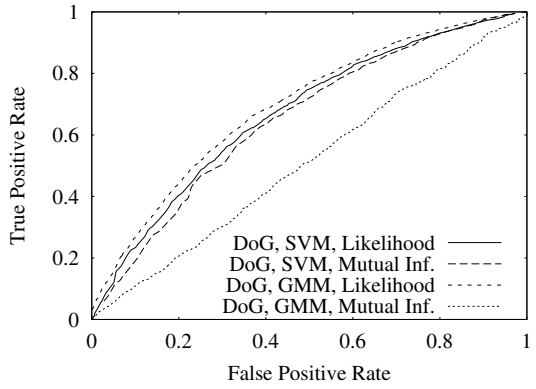
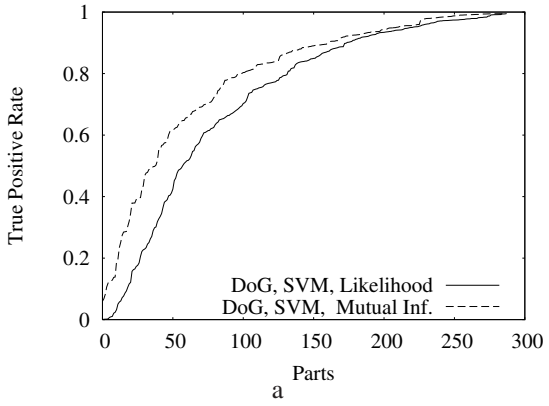


Figure 4. Comparison of the performance of the likelihood ratio and the mutual information for SVM and GMM. Regions are extracted with the DoG detector.

Descriptors. We have also compared the performance of the two detectors DoG and Harris-Laplace. Figure 6 shows the results for Harris-Laplace. We can observe that the ranking of the different combinations of classifier and feature selection techniques are the same as for DoG. Furthermore, Harris-Laplace and DoG show a similar performance. However, we have noticed that the behavior depends on the test image. Furthermore, Harris-Laplace detects less points on the background and therefore detects more true positives than DoG for a fixed number of false positives.

4.3. Car Recognition/Detection

In this section we illustrate the performance of our approach with two examples. Figure 7 shows results for DoG and Harris-Laplace as well as the two classification techniques. The first column displays the detected regions. The



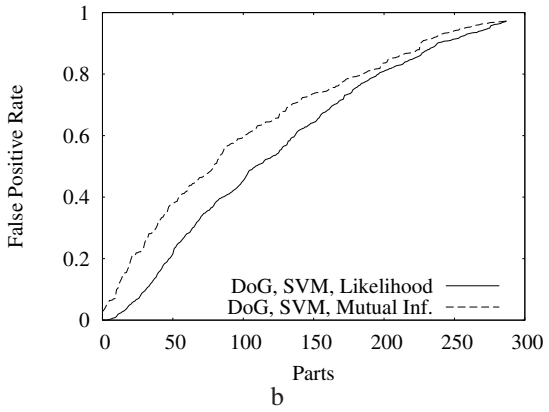


Figure 5. Comparison of the performance of the likelihood ratio and the mutual information for SVM. Regions are extracted with the DoG detector.

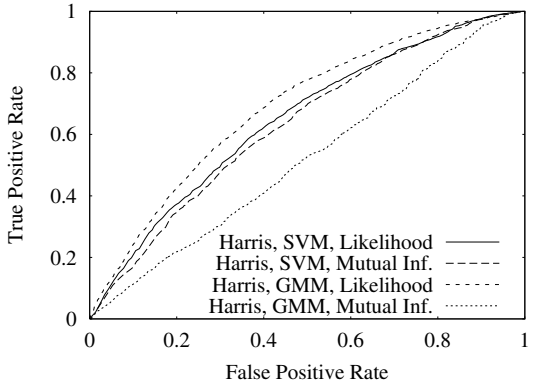


Figure 6. Comparison of the performance of the likelihood ratio and the mutual information for SVM and GMM. Regions are extracted with the Harris-Laplace detector.

second column shows the results of the 25 best parts obtained with SVM and selected by the likelihood ratio. The third column displays the results of the 100 best parts for GMM and likelihood ratio. We can see that the method allows to select car features. It can be further improved by adding relaxation.

Relaxation. The descriptors selected in Figure 7 are sparsely distributed over the object (car). We would like to obtain a dense feature map which permits segmentation of the object.

Given the selected features, we can use the order of selection to assign a probability to each descriptor. A descriptor which is classified by a more discriminative feature is assigned a higher probability. We can then use relaxation [13] to improve the classification of the descriptors. Relaxation reinforces or weakens the probabilities depending on the probabilities of the nearest neighbors (5 in our experiments). Figure 8 shows the descriptors classified as car features after applying the relaxation algorithm. Initial results based only on feature selection are shown in Figure 7 (DoG, SVM and likelihood). Compared to these initial results, we can clearly observe that more features are detected on the cars and less on the background, that is the overall performance is significantly improved. Further improvement is possible by integrating spatial constraints into the neighborhood relations of the relaxation process.

5. Conclusion and Future work

In this paper, we have introduced a method for constructing object-part classifiers and selecting the most discriminant ones. Object-parts are invariant to scale and rotation as well as illumination changes. Alignment of the training and test images is therefore not necessary. This paper has also illustrated the importance of feature selection and has compared different techniques. This comparison shows that likelihood is well suited for object recognition and mutual information for focus of attention mechanisms, that is rapid localization based on a few classifiers.

Learning of the parts is unsupervised, but the descriptors are manually marked as positive and negative. We plan to extend the approach to the weakly supervised case where the descriptors are unlabeled and only the images are marked as positive or negative. This should be straightforward in the case of classification with a Gaussian mixture model.

Acknowledgments

This work was supported by the European project LAVA. We thank S. Agarwal for providing the car images.

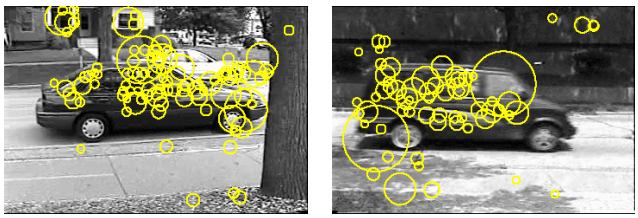


Figure 8. Improved results for object detection by adding relaxation.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, 2002.

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

[3] X. Chen, L. Gu, S. Li, and H.-J. Zhang. Learning representative local features for face detection. In *CVPR*, 2001.

[4] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.

[5] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods. In *CVPR*, 2003.

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

[7] S. Mahamud and M. Hebert. The optimal distance measure for object detection. In *CVPR*, 2003.

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

[9] C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38(1):15–33, 2000.

[10] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, 1991.

[11] R. Fergus, P. Perona and A.Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

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

[13] A. Rosenfeld, R. Hummel, and S. Zucker. Scene labeling by relaxation operations. IEEE *Transactions on Systems, Man and Cybernetics*, 6:420–433, 1976.

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

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

[16] S. Ullman, E. Sali, and M. Vidal-Naquet. A fragment-based approach to object representation and classification. In *4th International Workshop on Visual Form, Capri, Italy*, 2001.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume I, pages 511–518, 2001.

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

Figure 7. Results for two test images. The left column shows the interest regions detected with DoG and Harris-Laplace. The middle column displays the descriptors classified by the 25 best SVM classifiers selected with the likelihood ratio. The right column shows the results for the 100 best GMM classifiers selected with likelihood ratio.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE