Local Features for Object Class Recognition

Krystian Mikolajczyk, Bastian Leibe, and Bernt Schiele
Multimodal Interactive Systems
TU Darmstadt, Germany
kma,leibe,schiele@mis.tu-darmstadt.de

Abstract

In this paper we compare the performance of local detectors and descriptors in the context of object class recognition. Recently, many detectors / descriptors have been evaluated in the context of matching as well as invariance to viewpoint changes [20]. However, it is unclear if these results can be generalized to categorization problems, which require different properties of features. We evaluate 5 state-of-the-art scale invariant region detectors and 5 descriptors. Local features are computed for 20 object classes and clustered using hierarchical agglomerative clustering. We measure the quality of appearance clusters and location distributions using entropy as well as precision. We also measure how the clusters generalize from training set to novel test data. Our results indicate that extended SIFT descriptors [22] computed on Hessian-Laplace [20] regions perform best. Second score is obtained by Salient regions [11]. The results also show that these two detectors provide complementary features. The new detectors/descriptors significantly improve the performance of a state-of-the art recognition approach [16] in pedestrian detection task.

1. Introduction

Local photometric descriptors computed for interest regions have proved to be very successful in applications such as wide baseline matching [28], viewpoint invariant object recognition [9, 18], texture recognition [15], video data mining [26], image retrieval, robot localization and also in the recognition of object categories [7, 8, 23]. Consequently, various invariant detectors and descriptors have been proposed and evaluated in the literature [21, 22] in the context of viewpoint invariant matching.

It is unclear however, if such evaluations generalize to the challenging task of object class recognition. The employed evaluation criteria such as repeatability, precision, and recall are often not well-defined in the context of object class recognition, since – in general – there is no simple transformation relating instances within the same object class. Furthermore, the goal and requirements for object class recognition are different. For example, features should generalize beyond individual class members to enable the learning of a general object class model, and learning should be feasible from a small number of samples.

A possible way to evaluate different features is to use them within the context of various recognition approaches. However, many state-of-the-art approaches for object class recognition use clustering of local features as an intermediate level of representation [2, 3, 16, 26, 31]. The main motivations for this are the above-mentioned generalization and learning requirements for object class recognition. When aiming at evaluation of various detectors and descriptors, it is reasonable to start at this intermediate level of representation and to define evaluation criteria for feature clusters, rather than for individual features. However, one should then verify that the obtained results do indeed generalize to various recognition approaches.

The first major contribution of this paper are new criteria for evaluating feature detectors and descriptors in the context of object class recognition. More specifically, we require that feature clusters have high precision, i.e. that a cluster is representative for one class only. Although it is possible to make use of clusters that are shared by several classes [27], additional methods are necessary to resolve the ambiguities. In addition, we look for compact spatial location distributions of features (i.e. visually similar features should occur approximately at the same location on the object), as this property is essential for object localization.

The second major contribution is the evaluation of various state-of-the-art feature detectors and descriptors for 20 object classes from the CalTech 101 database. High performance is demonstrated by the Hessian-Laplace [20] and Salient region detectors [11]. The ranking of the top-performing detectors is different than presented in [21]. In particular, the MSER detector [19] obtains low scores, contrary to its performance for matching [21]. Gradient location and orientation histogram (*GLOH*) [22], which is an extension of the SIFT descriptor, is shown to outperform

SIFT, as well as the other descriptors. Furthermore, a paired t-test shows that the results are significant at a high confidence level.

Finally, as our third contribution, we improve a state-of-the art recognition system and validate the feature ranking on a challenging pedestrian detection task.

1.1. Related Work

Due to a large number of methods developed for similar computer vision problems, performance evaluation has gained more and more importance [6]. In the context of matching and viewpoint invariance, extensive evaluations of feature detectors [12, 20, 21, 25] and descriptors [22] are available. Performance is measured by the percentage of features simultaneously present in two images. Repeatability, precision and recall are used to evaluate descriptors. Several authors evaluate their descriptors in the context of matching [13, 18] or texture classification [15, 24] using different evaluation criteria and test data. However, the results cannot be directly compared or generalized to a class recognition problem.

Very little work has been done on the evaluation of features in the context of object class recognition. Performance of global descriptors was compared in [17] for recognition and image retrieval using nearest-neighbor matching and measuring the percentage of correct class label assignments. A small set of region detectors were evaluated for object category recognition in [12]. Manually annotated images were used to find correct correspondences. This solution is not practical for a large number of images representing different categories. Performance for object class recognition approaches is often reported for entire methods [4, 8, 16]. Although the same test data is used, it is unclear how much improvement is given by using different features.

Related work on evaluation of clustering can also be found in document classification domain [14]. In this context, entropy and intersections between clusters and classes are frequently used to evaluate the quality of clusters.

1.2. Overview

Section 2 briefly describes the detectors and descriptors used in our comparison. In section 3 we explain our experimental setup and the feature cluster representation. Section 4 introduces our novel evaluation criteria for object class recognition and the dataset. In section 5 we discuss the experimental results.

2. Local Features

Many different techniques for detecting and describing local image regions have been developed. In this section we briefly describe the detectors and descriptors used in our evaluation. See [20, 21, 22] for a survey on invariant detectors, descriptors and implementation details. Original implementations for all detectors and descriptors used in this paper are also available [1].

2.1. Region Detectors

Region detectors use different image measurements and can be invariant to various transformations. In this paper we focus on five different scale invariant detectors. There is a number of detectors invariant to affine transformations which provide elliptical regions. However, the region locations and scales are the same as in their scale invariant versions only the shape of regions varies.

The detectors provide regions which are used to compute descriptors. In this evaluation we use five detectors :
Harris-Laplace regions [20] are detected by the scale-adapted Harris function and selected in scale-space by the Laplacian-of-Gaussian operator. Harris-Laplace detects corner-like structures.
DoG regions [18], are localized at local scale-space maxima of the difference-of-Gaussian. This detector is suitable for finding blob-like structures.
Hessian-Laplace regions [21] are localized in space at the local maxima of the Hessian determinant and in scale at the local maxima of the Laplacian-of-Gaussian.
Salient regions [11] are detected in scale-space at local maxima of the entropy. The entropy of pixel intensity histograms is measured for circular regions of various size at each image position. These regions were successfully used in object class recognition [8].
Maximally Stable Extremal Regions, (MSER) [19] are components of connected pixels in a thresholded image. A watershed-like segmentation algorithm is applied to image intensities and segment boundaries which are stable over a wide range of thresholds define the region. To obtain the position of the regions we compute the average x and y pixels locations. The size is given by a geometric mean of the eigenvalues of the second order moments matrix, computed for the pixel locations.

Region normalization The detectors provide circular regions the size of which depends on the detection scale. All the regions are mapped to a circular region of constant radius to obtain scale invariance. According to [22] the size of the normalized region is arbitrarily set to 41 pixels.

2.2. Descriptors

In the following we present the descriptors used in our experimental evaluation. We selected a subset of descriptors evaluated in [22] which showed superior performance in the context of matching images with viewpoint changes.
SIFT descriptors [18] are 3D histograms of gradient locations and orientations, where locations are quantized into a

4x4 location grid and the gradient angle is quantized into 8 orientations. The resulting descriptor is of dimension 128.

Gradient location-orientation histogram (GLOH) [22] is an extension of the SIFT descriptor designed to increase its robustness and distinctiveness. Compared to SIFT, the histogram is computed for 17 location and 16 orientation bins in a log-polar location grid. PCA is used to reduce the dimension to 128.

PCA-SIFT [13] descriptor is a vector of image gradients in x and y direction computed within the support region. The dimension is reduced to 36 with PCA.

Moment invariants [30] are computed up to 2^{nd} order and 2^{nd} degree for derivatives of an image patch: $M^a_{pq} = \frac{1}{xy}\sum_{x,y} x^p y^q [I_d(x,y)]^a$, of order $p+q$ and degree. I_d is the image gradient in direction $d = x, y$. This results in a 20-dimensional descriptor.

Cross correlation (CC). To obtain this descriptor the region is smoothed, uniformly sampled at 9x9 pixel locations, and normalized with its mean and standard deviation.

All descriptors are normalized with a covariance matrix. Thus, Euclidean distance can be used to compute the similarity between two descriptors.

3. Feature Cluster Representation

In the following we describe our clustering approach and object class representation.

3.1. Clustering

For our evaluation we build clusters using average-link agglomerative clustering. This clustering method does not depend on initialization, unlike partitional clustering such as K-means or EM-clustering. Moreover, it was indicated superior to K-means in [10]. Given F features computed for images of all object categories the clustering is initialized with F clusters each containing 1 feature only. At each iteration the two most similar clusters are merged. The similarity between two clusters is an average distance between all features f in these two clusters:

$$\frac{1}{NM}\sum_{m}^{M}\sum_{n}^{N}(f_{km}-f_{ln})^2 = \sigma_k^2+\sigma_l^2+(\mu_k-\mu_l)^2 \leq v \quad (1)$$

where N and M are numbers of features in clusters k and l; μ_k and μ_l are the cluster centers; σ_k^2 and σ_l^2 are the variances. The agglomerative clustering produces a hierarchy of merging steps up to the point where the cut-off criterion stops the clustering process. We thus obtain clusters for which the similarity distance between every pair of clusters is above v.

The standard algorithm for average-link clustering, as found in most text books, requires the computation of an $(\#F)^2$ similarity matrix, which makes it inapplicable for clustering more than 20 000 features. Our evaluation data consists of more than 100 000 features, therefore we use an extension of the Reciprocal Nearest Neighbor (RNN) algorithm [5], which reduces quadratic space complexity to linear.

3.2. Class Representation

Features, which are computed on the training data of various classes, are grouped in appearance clusters with spatial location distributions. An appearance cluster is represented by a mean vector of all descriptor vectors in the cluster. A cluster can contain features from several classes and corresponding location distributions. A spatial distribution is a quantized histogram of feature locations, at which features occur on a given object class within one cluster. We use a 5x5x4 location grid for x, y, and $scale$ dimensions. The histograms are normalized such that they reflect probabilities of spatial locations for the appearance clusters. A spatial location distribution is estimated for each appearance cluster from all features that match to this cluster. Figure 1 illustrates the appearance clusters and location distributions.

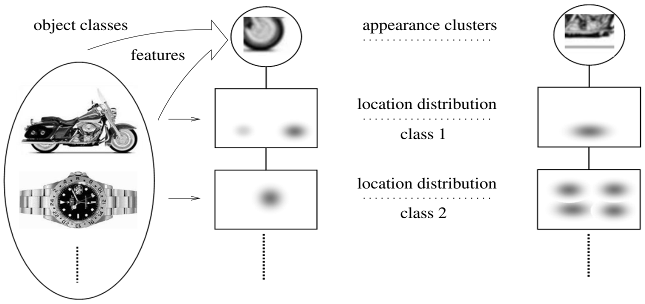


Figure 1. Feature cluster representation.

4. Feature Evaluation Criteria

In this section we define novel evaluation criteria applicable in the context of object class recognition. We introduce several measures which emphasize different properties of features. We also present the data set used in our tests.

4.1. Appearance Clusters

We evaluate the clustering properties of features with average cluster precision. This criterion measures how the clusters are shared between different classes. To compute an average precision for a class we take into account only clusters in which the class dominates, otherwise the precision would decrease significantly due to a large number of clusters which contain only one feature of that class. Suppose there are M clusters in which object class a dominates. Average precision P_{Ca} for these clusters is then defined by

$$P_{Ca} = \frac{1}{M}\sum_{j=1..M} p_{j_a} \quad (2)$$

where p_{j_a} is a probability of class a in cluster j illustrated in figure 2. The precision attains maximum if each cluster contains features from one class only.

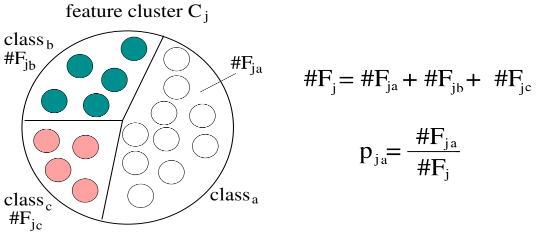


Figure 2. Cluster precision p_{ja}. $\#F_j$ is the number of features in cluster j; $\#F_{ja}$ is the number of features from class a in cluster j.

To evaluate generalization properties of features we apply a criterion which measures the precision of test features matched to clusters. We compute features for test images and match them to the clusters built from the training data. A feature matches to a cluster if the distance to the cluster center is below a threshold. For a given number of clusters, the similarity threshold is the minimum distance between two clusters defined in equation 1. We define matching precision for test features of class a as:

$$P_{Ma} = \frac{1}{J} \sum_j^J p_{ja} \qquad (3)$$

where p_{ja} is defined in figure 2. One feature can match to several clusters and J is the total number of matches.

4.2. Location Distributions

In the following we propose a measure to evaluate location distributions of clusters. We use entropy to measure the compactness of the distributions. For a given class the entropy is defined by:

$$E_{loc} = \frac{1}{C} \sum_c^C \sum_{\mathbf{x}}^X -p_c(\mathbf{x}) \log\left(p_c(\mathbf{x})\right) \qquad (4)$$

where $p_c(\mathbf{x})$ is a probability of spatial location $\mathbf{x}$ for appearance cluster c. X is the number of bins in the quantized location histograms and C is the total number of clusters. Ideally, a cluster will have compact location distributions, that is the probability of a feature that matches to the cluster is high at few locations and very low in other positions.

4.3. Detectors Complementarity

To find complementary feature detectors we measure how often two types of regions occur in the same clusters, that is how often they detect similar local structures. Given clusters with features extracted by detectors e and d from various object classes we estimate a complementarity score:

$$complementarity = 1 - 2 \cdot \frac{1}{C} \sum_c^C \min(p_{jd}, p_{je}) \qquad (5)$$

where $p_{jd},\ p_{je}$ are probabilities of feature type d and e in cluster j defined in a similar manner to class probability in figure 2. Constant values in the equation normalize the score to the range $[0..1]$. Complementarity is 1 if no cluster is shared between two feature types.

4.4. Test Data

There are three data sets used in our tests. The features are evaluated on real images of the first 20 object classes (alphabetical order) from the Caltech 101 categories. As the Caltech 101 database contains only 30 images for some categories, we limit our training set to 20 images per category (400 images) and the test set to 10 images per category (200 images). The number of features provided by different detectors with default parameter settings vary from 50000 to 100000. To evaluate the localization properties of features we use three object categories, namely faces, cars, and motorbikes, since the training images are roughly aligned only for these objects. We use 100 images for training and 100 for testing. The same 3 categories are used for testing the complementarity of different detectors. In all data sets the test images are different from the training images. High recognition performance on the Caltech 101 categories was already reported in [4, 8, 29]. We therefore use more challenging test data with pedestrian images to validate the evaluation results. The training set consists of 105 pedestrian images. The test set contains 209 images with 595 pedestrians of different sizes, with occlusion and background clutter. A few examples are displayed in figure 4.

5. Experimental results

In this section we present and discuss the results obtained for the evaluation criteria proposed in this paper. In section 5.1 we evaluate the quality of appearance clusters and precision of test features which match to the clusters. We also perform a paired t-test to obtain the confidence levels for the results. In section 5.2 we discuss the evaluation results for spatial location distributions. In section 5.3 we investigate the complementarity of different detectors and in section 5.4 we validate the evaluation results with object recognition test. There are 5 detectors and 5 descriptors, which make 25 combinations. We therefore show the results only for selected pairs. To compare the performance of different detectors we combine them with the same top-performing descriptor (GLOH). Similarly, the presented results for different descriptors are obtained by using the same top-performing detector (Hessian-Laplace). The ranking of detectors remains similar regardless of the descriptor we use, similarly the ranking for different descriptors is independent of the detector, only absolute scores differ.

5.1. Appearance Clusters

The evaluation carried out in this section is done for 20 object classes.

Feature density. The agglomerative clustering approach can produce a number of clusters, which can vary from one up to a level at which each cluster contains only one feature. For example, if one detector provides less features than the other and we build the same number of clusters for both, then in the first case there will be more clusters containing a single feature. The evaluation would be biased by the number of features and the number of clusters. To make the results comparable we refer to average density of features per cluster. We have chosen experimentally a range of densities for which the regions within a cluster remain visually similar. We first measure the percentage of clusters which contain only one feature. Typically, single member clusters generalize poorly since the local feature was found only once in the entire training data. Figures 3(a) and 3(b) show the ratio of single member clusters to the total number of clusters. Figure 3(a) shows the ratio for different detectors. The lowest number of single member clusters is obtained with Hessian-Laplace, Harris-Laplace, and Salient regions detector. MSER detector provides very discriminant regions which do not cluster well. Figure 3(b) shows the results for different descriptors and the worst score is obtained by moments and cross-correlation of patches.

Cluster precision. To evaluate the quality of clusters we compute average cluster precision defined by equation 2. The higher the precision, the less clusters are shared between different classes. Figure 3(d) displays the results for different detectors and figure 3(g) for different descriptors. The highest precision is obtained by Hessian-Laplace and Salient region detector. GLOH descriptor gives the highest score. SIFT descriptor obtains high precision for small density of features per cluster but the precision drops down as the density increases. This indicates that the descriptor is very discriminant but less tolerant to appearance variations.

Matching score. In the following we evaluate matching precision of test features to clusters. The matching precision is defined in equation 3. The similarity threshold is the closest distance between two clusters in the space. We can use different similarity threshold to match features to clusters i.e. divide the clustering thresholds by a constant factor. Figure 3(e) shows the matching precision for different detectors and figure 3(h) for different descriptors with fixed density of 10 features per cluster and varying threshold. We observe that the precision is low and very similar for different detectors and descriptors if we directly apply the clustering threshold. The precision increases if this threshold is reduced by a factor of 2. Figure 3(c) shows the ratio of test features which can still be matched to clusters with a given threshold. Approximately 90% of features can be matched with the clustering threshold reduced by a factor of 2. We therefore use this threshold for further experiments. Figure 3(f) displays matching precision with respect to the feature density for different detectors and figure 3(i) for different descriptors. The best score is obtained by Hessian-Laplace regions and GLOH descriptor. MSER regions and moments obtain low score.

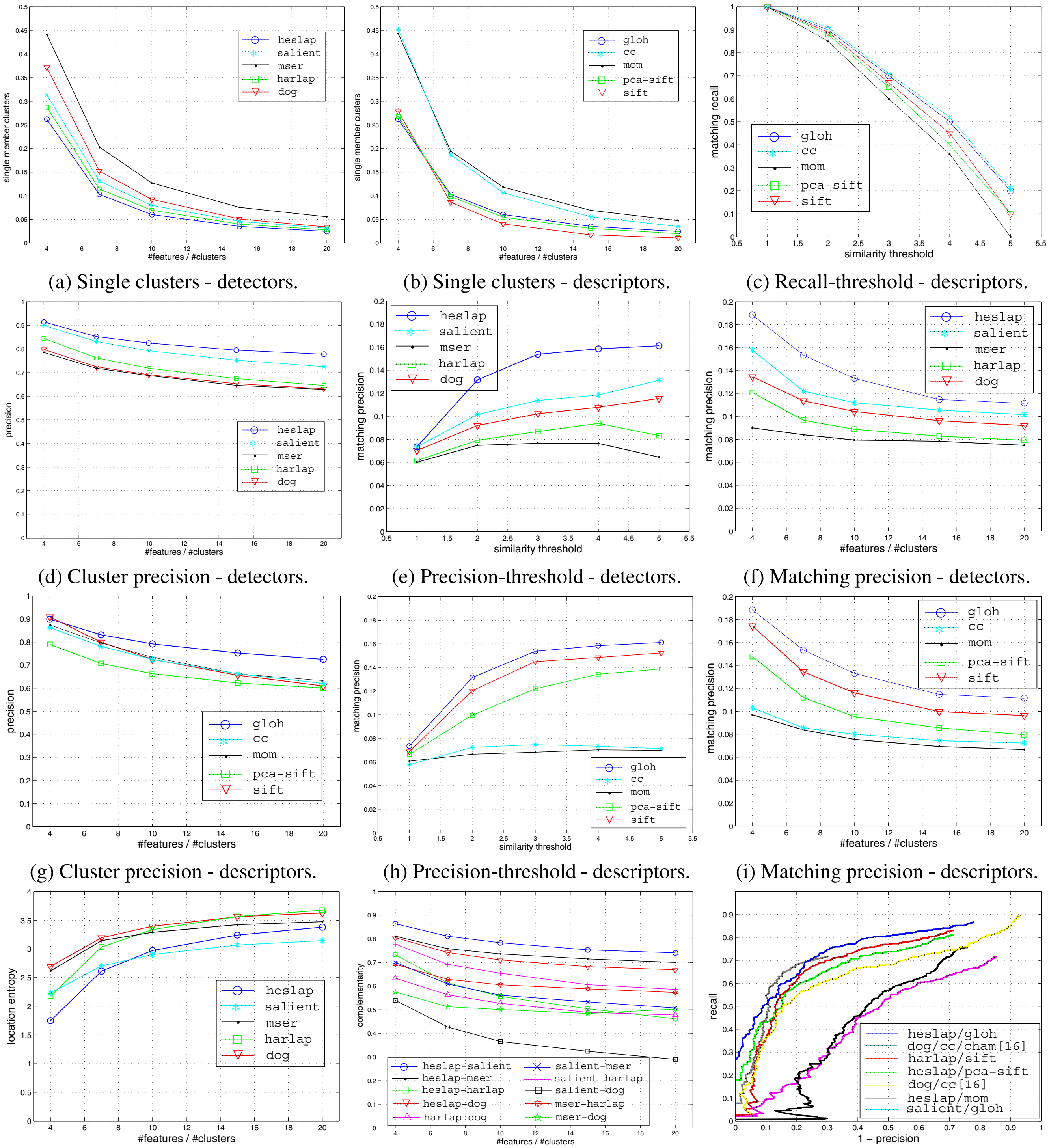
Paired t-test. The results presented in figure 3(a)-(i) are averaged over 20 object classes (unless stated otherwise), that is each point on a curve is a mean score for all classes. The average values can be dominated by one class only if the score for this class is significantly larger than for all other classes. To show that the improvement is significant for all classes we perform paired t-test. The test is done for each pair of points on two curves at the corresponding density $\#features/\#clusters$. For a given degree of freedom, which is 20 in our case, the t-value indicates the probability that the difference between scores for different features is statistically insignificant across all object classes. Table 1 shows the minimum confidence levels for arbitrarily chosen pairs. For example, the confidence level of 0.01 for Hessian-Laplace with GLOH descriptor (hes-gloh) and DoG with PCA-SIFT (dog-pca) indicates that the probability of making improvement by using the first combination is at least 0.99. According to paired t-test most of the results are significant at a high confidence level.

	hes	sal	dog	har	mser
	gloh	gloh	gloh	gloh	gloh
hes-sift	-	-	0.01	0.01	0.01
sal-gloh	-	-	0.01	0.01	0.01
sal-mom	-	-	0.05	0.05	0.05
hes-pca	0.05	-	-	0.05	0.01
mser-cc	0.01	0.01	0.01	0.01	0.05
har-gloh	0.01	0.01	0.05	-	0.01
dog-pca	0.01	0.01	0.05	0.05	-

Table 1. Paired t-test. Maximum probability that the difference between feature performance is insignificant. "-" if the probability is higher than 0.05. hes - Hessian-Laplace; har - Harris-Laplace; sal - salient regions detector; pca - PCA-SIFT; cc - cross-correlation of patches.

5.2. Localization

In this section we present the evaluation results of feature localization properties. The results are presented only for detectors, since localization accuracy is a property of region detectors. Test images are roughly aligned, that is similar object parts occur at the same locations and scales. We build appearance clusters and spatial location distributions for each detector by matching test features to the clusters, as described in section 3.1. We then compute the entropy of



(a) Single clusters - detectors.

(b) Single clusters - descriptors.

(c) Recall-threshold - descriptors.

(d) Cluster precision - detectors.

(e) Precision-threshold - detectors.

(f) Matching precision - detectors.

(g) Cluster precision - descriptors.

(h) Precision-threshold - descriptors.

(i) Matching precision - descriptors.

(j) Localization

(k) Complementarity

(l) Pedestrian detection.

Figure 3. Feature evaluation results. (a) Ratio of single member clusters for detectors. (b) Ratio of single member clusters for descriptors. (c) Recall with respect to similarity thresholds for descriptors. (d) Cluster precision for detectors. (e) Matching precision with respect to similarity thresholds for detectors. (f) Matching precision with respect to feature density for detectors. (g) Cluster precision for descriptors. (h) Matching precision with respect to similarity thresholds for descriptors. (i) Matching precision with respect to feature density for descriptors. (j) Entropy of location distributions. (k) Complementarity score. (l) Precision-recall for pedestrian detection.

the distributions (cf. equation 4). Figure 3(j) shows the results for different detectors combined with GLOH. Location distributions computed for Hessian-Laplace and Salient regions have the lowest entropy. This means that for an average appearance cluster, the probability of spatial occurrence on the object is concentrated around few locations.

5.3. Complementarity

Complementarity is estimated for different pairs of detectors. For a given pair of detectors we extract regions from training images, compute GLOH descriptors and cluster them together. One cluster space is built for each pair of detectors. As in the localization test (cf. section 5.2) we use the test data with 3 object categories. The score is measured according to equation 5. Two detectors obtain high score if their regions are dissimilar. Figure 3(k) shows the results where each curve corresponds to one pair of detectors. Hessian-Laplace and Salient is the most complementary pair of all evaluated detectors. The score is approximately 0.8, which means that only 10% of clusters is shared (cf. equation 5). These detectors find different image structures which form separate clusters. The next complementary pair is Hessian-Laplace and MSER. Low score is obtained by Salient region detector and DoG. These results are surprising, since DoG and Hessian-Laplace are based on similar filters (LoG and DoG). However, feature locations both in scale and image plane differ due to different functions, which are used to evaluate filter responses.

5.4. Object Class Recognition

To show that the results reported in this paper hold for different test data and a different object class we apply a subset of detectors/descriptors to a challenging pedestrian detection problem. We apply recognition approach of [16] using the original code with new features. A pedestrian model is trained from 105 images and represented by local appearance clusters with spatial location distributions. The density of $\#features/\#clusters$ is approximately 5. We train one model for each combination detector / descriptors and apply the detector. The test data is described in section 4.4 and examples of images are displayed in figure 4. Note the difficulty of the recognition task in these images. A detection is correct if the intersection to the union of the detection and the ground truth bounding boxes is at least 0.5. Evaluation criterion is precision-recall. Figure 3(l) shows the detection results. Hessian-Laplace and GLOH descriptor improve the detection score at EER by 10% compared to the one reported in [16](`dog/cc`). For comparison we also show the results for the complete recognition system which applies additional verification stage based on chamfer distance (`dog/cc/cham` [16]). For high precision this score is also lower than Hessian-Laplace/GLOH. Hessian-Laplace is closely followed by Harris-Laplace detector. This detector provides fewer regions than Hessian-Laplace but they are mostly localized on the pedestrians. Salient regions combined with (GLOH) obtain low score. This method extracts large regions which also cover the background, therefore the performance is lower in cluttered images. Hessian-Laplace with moments obtain very low score. These results also show that careful selection of both detector and descriptors is important. Results for other combinations are consistent with the evaluation presented in this paper. Figure 4 shows some examples of pedestrian images with correct detections obtained with the Hessian-Laplace detector and GLOH descriptors.

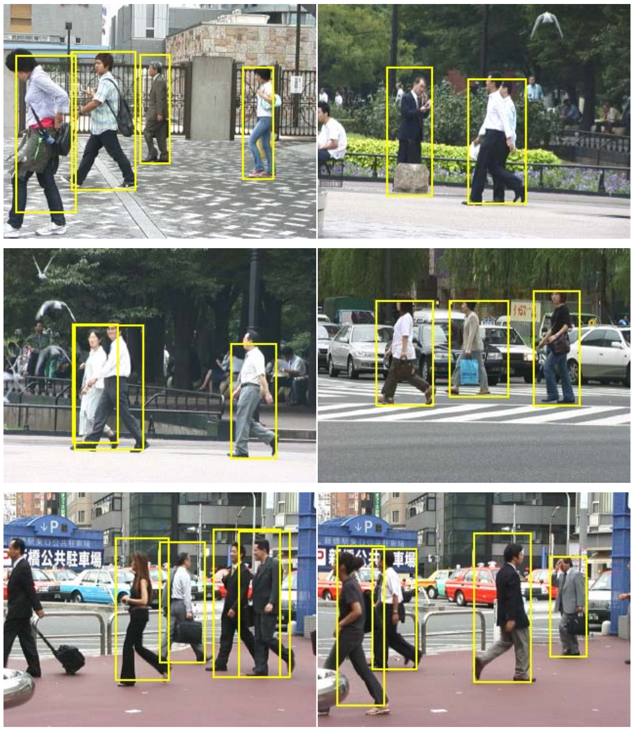


Figure 4. Examples of pedestrian detections obtained with Hessian-Laplace detector and GLOH descriptors.

Summary and Conclusions

In this paper we have presented an experimental evaluation of local features in the context of recognition and classification of object categories. We have proposed several criteria to evaluate different properties of features. We have compared the performance of detectors and descriptors computed with recently proposed methods. Finally, we significantly improve the performance of the recognition system [16] with new features.

In the presented evaluation the GLOH descriptor computed on Hessian-Laplace regions systematically obtains the highest score. Salient regions also perform well. It is important to note that these two detectors obtain the highest complementarity score. MSER detector and PCA-SIFT descriptor seems to be more suitable for matching [21] than for recognition of categories used in this evaluation. The MSER detector provides very distinctive regions but too few for reliable recognition. High performance of the extended SIFT descriptor (GLOH) follows the results obtained in the context of matching [22], which confirms the robustness and the distinctive character of the region-based SIFT descriptors. A paired t-test shows that the results are significant for all evaluated categories. Furthermore, the results are validated in the context of a complete state-of-the-art recognition system on independent data set. We did not find any evidence that there is a detector or a descriptor which is more suitable for one particular category of objects. Additional experiments have to be carried out to clarify that. We currently investigate other clustering schemes, e.g. partitional clustering, and evaluate the performance of the recognition approach using combinations of different features.

Acknowledgments

This work has been funded by the EU project CoSy (IST-2002-004250).

References

[1] www.robots.ox.ac.uk/~vgg/research/affine

[2] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, 2004.

[3] M. Burl, M. Weber, and P. Perona A Probabilistic Approach to Object Recognition Using Local Photometry and Global Geometry. In *ECCV*, pages 628-0641, 1998.

[4] T. Berg, A. Berg, and J. Malik. Shape Matching and Object Recognition using Low Distortion Correspondence. In *CVPR*, pages 26–33, 2005.

[5] J.P. Benzécri. Construction d'une Classification Ascendante Hiérarchique par la Recherche en Chaîne des Voisins Réciproques. *CAD*, 7(2):209–218, 1982.

[6] H. I. Christensen and P. J. Phillips, editors. *Empirical Evaluation Methods in Computer Vision*. World Scientific Publishing Co., 2002.

[7] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, pages 634–640, 2003.

[8] R. Fergus, P. Perona, and A.Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271, 2003.

[9] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *ECCV*, pages 40–54, 2004.

[10] A.K. Jain and R.C. Dubes. Algorithms for Clustering Data, Prentice-Hall, 1988

[11] T. Kadir, M. Brady. Scale, Saliency and Image Description. *IJCV*, 45(2):83–105, 2001.

[12] T. Kadir, A. Zisserman and M. Brady. An affine invariant salient region detector. *ECCV*, pages 404–416, 2004.

[13] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *CVPR*, pages 511–517, 2004.

[14] G. Kowalski. Information Retrieval Systems - Theory and Implementation. Kluwer Academic Publishers, 1997.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods. In *CVPR*, pages 319–324, 2003.

[16] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, pages 878–885, 2005.

[17] B. Leibe and B. Schiele. Analyzing Appearance and Contour Based Methods for Object Categorization. In *CVPR* , pages 409–415, 2003

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004.

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002.

[20] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *IJCV*, 2005.

[22] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):31–47,2005.

[23] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV*, pages 71–84, 2004.

[24] T. Randen and J. H. Husoy. Filtering for texture classification : A comparative study. *PAMI*, 21(4):291–310, 1999.

[25] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *IJCV*, 37(2):151–172, 2000.

[26] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*. pages 1470–1478, 2003.

[27] A. Torralba, K. Murphy, and W. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*,pages 762–769, 2004.

[28] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 1(59):61–85, 2004.

[29] J. Thureson and S. Carlsson. Appearance Based Qualitative Image Description for Object Class Recognition. *ECCV*, pages 518-529, 2004.

[30] L. Van Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *ECCV*, pages 642–651, 1996.

[31] M. Weber, M. Welling, and P. Perona Unsupervised Learning of Models for Recognition. In *ECCV*, pages 628-0641, 2000.