Segmentation Based Multi-Cue Integration for Object Detection

Bastian Leibe
ETH Zurich
Zurich, Switzerland

Krystian Mikolajczyk
University of Surrey
Guildford, UK

Bernt Schiele
TU Darmstadt
Darmstadt, Germany

Abstract

This paper proposes a novel method for integrating multiple local cues, i.e. local region detectors as well as descriptors, in the context of object detection. Rather than to fuse the outputs of several distinct classifiers in a fixed setup, our approach implements a highly flexible combination scheme, where the contributions of all individual cues are flexibly recombined depending on their explanatory power for each new test image. The key idea behind our approach is to integrate the cues over an estimated top-down segmentation, which allows to quantify how much each of them contributed to the object hypothesis. By combining those contributions on a per-pixel level, our approach ensures that each cue only contributes to object regions for which it is confident and that potential correlations between cues are effectively factored out. Experimental results on several benchmark data sets show that the proposed multi-cue combination scheme significantly increases detection performance compared to any of its constituent cues alone. Moreover, it provides an interesting evaluation tool to analyze the complementarity of local feature detectors and descriptors.

1 Introduction

Local feature based approaches have shown considerable promise for dealing with the large degree of intra-category variation and partial occlusion inherent in real-world categorization and detection tasks. Consequently, many approaches have been developed that use local features in different ways [1, 6, 4, 10, 12], and considerable progress has been made in the design and understanding of the underlying feature detectors and descriptors [12, 14]. Yet, each feature descriptor and detector can only capture part of the information contained in the image, and indeed its value for an application depends on the degree to which it can distill exactly the right kind of information for a specific purpose. As a consequence, the better a descriptor or detector is suited to a specific task, the more likely it is to degenerate when the task conditions deviate too far from its target scenario. In order to be both discriminative and robust, an application should therefore utilize a combination of different local cues.

Several recent studies have evaluated the suitability of various local features in the context of object identification [14] and categorization tasks [13]. However, those studies have only considered each cue in isolation. For multi-cue integration, it is also important to know how the different cues interact, i.e. how correlated their responses are and what new information an additional cue can contribute. However, this information is difficult to retrieve, as different cues are often not directly comparable, both because they typically have different dimensionalities and because they represent information in different ways.

Previous research has therefore mainly focused on *classifier combination*, i.e. on the problem of fusing the outputs of several "black-box" classifiers, possibly with associated confidence ratings [20, 9, 7, 15]. This approach is valid if the classifiers are independent. In our application, however, their outputs are often correlated, and the degree of correlation

may vary from image to image. Rather than just to fuse the outcomes of several classifiers, we therefore need to explore how the underlying information and the respective support in the image can be combined.

In this paper, we present a flexible integration scheme which combines different local cues in an opportunistic manner depending on their explanatory power for the image at hand. The integration proceeds in two steps. First, the sampled features are represented in terms of their similarity to a set of prototypes, an *appearance codebook*, which has been learned for each cue separately. Together with their learned spatial distributions, those codebook prototypes convert the activations from matching features into a probability distribution for possible object locations and scales. This makes the cues comparable. However, their individual responses might still be correlated. Therefore, the second step backprojects the extracted object hypotheses to the image in order to determine for each cue separately which image pixels were responsible for a detection and how much each pixel contributed to the cue's response. By comparing the overlap in their supporting area, our approach can determine the complementarity between two cues and integrate their contributions more robustly.

This paper makes the following three contributions. Firstly, it develops a robust multi-cue integration approach that can be applied regardless of whether the cues are correlated or not. The proposed scheme is directly interpretable and opens up interesting venues for analyzing the complementarity of local cues. Secondly, it presents an extensive evaluation of state-of-the-art region detectors and descriptors in the context of multi-cue integration. The obtained results allow us to rank the cues based on their individual performances and to formulate clear usage guidelines for their combination. Last but not least, experimental results on several challenging data sets show that the proposed multi-cue integration scheme increases object detection performance significantly. The improvement is particularly prominent for the detection precision and leads to high recognition rates at the zero-false-positive level. The paper is structured as follows. The next section discusses related work. Section 2 then reviews the basic recognition approach. Extending this approach, we derive our proposed multi-cue integration scheme in Section 3. Section 4 describes our experimental setup, and Section 5 finally presents the results of our evaluation.

Related Work. Many authors have stressed the need for integrating multiple global or local cues in order to increase robustness of recognition [18, 11, 7]. In practice, multi-cue systems for object recognition have often been implemented by combining classifiers [20, 9, 7] or by using cue confidences in a voting scheme [3, 15]. However, these approaches are often static in that they use a fixed confidence rating per cue, e.g. based on previously observed performance. As such, they cannot readily adapt to novel settings when a cue's performance characteristics degrade due to changed environmental conditions. It has therefore been argued that cue weights should be adapted dynamically [17]. For tracking scenarios, cue integration techniques have been proposed which combine cues probabilistically based on their estimated likelihood [19]. However, in the context of single-frame object detection, no such mechanism has been known. In this paper, we propose such a mechanism based on the top-down segmentation approach by [10].

2 Recognition Approach

Our multi-cue recognition approach closely builds upon the Implicit Shape Model (ISM) formalism by [10, 11], which combines object detection and top-down segmentation capabilities. This model represents an object category by a set of local appearance clusters (a *codebook*) and their spatial occurrence distributions. Since a basic knowledge of this approach is necessary to understand our method, we will briefly review its main components.

Training. For training, local features are extracted from the training images and clustered to form the codebook [1, 10]. In a second run over the training data, the spatial occurrence distributions are estimated by recording for each codebook entry all matching locations on the training objects. Together with each occurrence, the approach stores a local segmentation mask, which is later used for inferring top-down segmentations.

ISM Recognition. During recognition, local features are extracted from the image and matched to the codebook. Each matching codebook entry then casts votes for possible object locations and scales in a probabilistic extension of the Hough transform [10]. For each hypothesis, the approach then computes a top-down segmentation and finally selects the subset of hypotheses that best explain the image content under the constraint that each pixel can be assigned to at most one hypothesis.

3 Multi-Cue Integration

We now present our novel approach for integrating multiple local cues. In the context of this paper, we understand this as a combination of different local descriptors, but also of different region detectors, since their preference for certain image structures influences the characteristics of the sampled information. As already mentioned before, the question how to combine local cues has no obvious answer, since they are typically not directly comparable.

We therefore proceed in two stages. The first stage extends the recognition procedure to include multiple cues. Its main purpose is to express the cues on a common basis, so that their information can be pooled and initial object hypotheses can be found. This stage still ignores cue correlation. Indeed, it has no other choice, since correlation can only be measured relative to a reference hypothesis, and hypotheses are only available after the stage has been executed. However, the second stage then reveals the correlation by backprojecting hypotheses to the image and computing a top-down segmentation for each cue. This step extends the ISM segmentation algorithm to deal with multiple cues. The obtained segmentations show on a per-pixel level which image structures were responsible for a cue's response. The correlation between two cues can then be expressed as the overlap of their respective $p(\textit{figure})$ probability maps. Once the cue correlation has been identified, the next question is how to use this information to improve recognition performance. In the last part of this section, we present three combination criteria that relate to different strategies for this step.

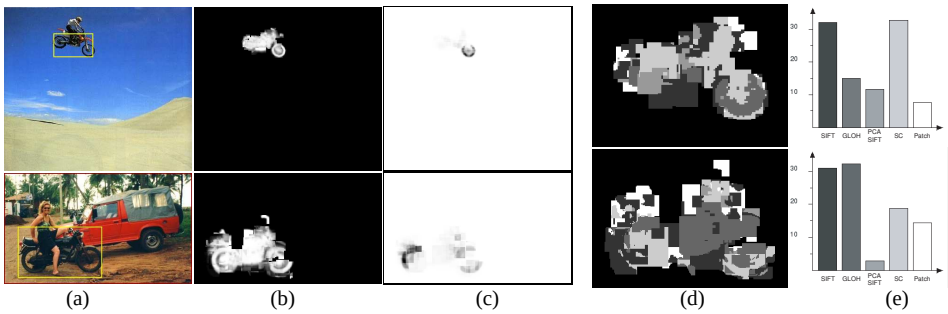
Initial Recognition Stage. The key to integrating multiple local cues is to express them on a common basis. We create such a basis by representing sampled features through their similarity to stored prototypes. We therefore extend the recognition approach by keeping a separate codebook $\mathcal{C}^q$ for every cue q. Let $\mathbf{e}$ be a local descriptor computed at location ℓ. When matched to the codebook, it may activate several codebook entries $\mathcal{C}_i^q$ with probabilities $p(\mathcal{C}_i^q|\mathbf{e})$. Each matched codebook entry then votes for instances of the object category o_n at different locations and scales $\lambda = (\lambda_x, \lambda_y, \lambda_\sigma)$ according to its learned occurrence distribution $P(o_n, \lambda|\mathcal{C}_i^q, \ell, q)$. A feature's contribution to an object hypothesis can thus be expressed as

$$p(o_n, \lambda|\mathbf{e}, \ell, q) = \sum_i P(o_n, \lambda|\mathcal{C}_i^q, \ell, q) p(\mathcal{C}_i^q|\mathbf{e}). \quad (1)$$

The contributions from all cues are pooled in a shared 3-dimensional voting space, from which maxima are extracted by Mean Shift Mode Estimation using a scale-adaptive kernel K [11], marginalizing over the cues q_m

$$\hat{p}(o_n, \lambda) = \frac{1}{nb(\lambda)^3} \sum_m \sum_k \sum_j p(o_n, \lambda_j|\mathbf{e}_k, \ell_k, q_m) K\left(\frac{\lambda - \lambda_j}{b(\lambda)}\right) p(\mathbf{e}_k, \ell_k|q_m) p(q_m), \quad (2)$$

where $b(\lambda)$ is the scale-adaptive kernel bandwidth; $p(\mathbf{e}_k, \ell_k|q_m)$ is an indicator variable specifying which image patches and locations have been sampled for q_m; and $p(q_m)$ is a prior



(a) (b) (c) (d) (e)

Figure 1: *Visualization of the multi-cue integration stages: (a) initial detection, (b) top-down segmentation, (c) $p(figure)$ maps obtained by average combination, (d) closeup view of the* argmax *visualization (cf. eq.(9)) , (e) histogram of relative cue contributions.*

determining how much this cue can be trusted. This prior can be set to reflect previously observed performance. In order to avoid any bias, however, we leave it at a uniform setting.

Multi-Cue Segmentation. Once a hypothesis $h = (o_n, \lambda)$ has been found, its top-down segmentation can be inferred by backprojecting the supporting votes to the image and combining them with the local patch segmentation masks $p(\mathbf{p} = fig. | o_n, \lambda, c_i^q, \ell)$ that have been stored for each recorded codebook occurrence during training. As shown in [10], the per-pixel probabilities of each pixel containing *figure* or *ground* can then be obtained by a double marginalization, first over sampled features, then over codebook entries. We adapt this formulation here to compute a separate segmentation for each cue

$$p(\mathbf{p} = fig. | o_n, \lambda, q) = \sum_{\mathbf{p} \in (\mathbf{e}, \ell)} \sum_i p(\mathbf{p} = fig. | o_n, \lambda, \mathbf{e}, c_i^q, \ell, q) p(\mathbf{e}, c_i^q, \ell, q | o_n, \lambda) \quad (3)$$

$$= \sum_{\mathbf{p} \in (\mathbf{e}, \ell)} \sum_i p(\mathbf{p} = fig. | o_n, \lambda, c_i^q, \ell) \frac{p(o_n, \lambda | c_i^q, \ell, q) p(c_i^q | \mathbf{e}) p(\mathbf{e}, \ell)}{p(o_n, \lambda)} \quad (4)$$

Based on these results, the final segmentation is computed by building the likelihood ratio between *figure* and *ground* probabilities.

Segmentation-Based Cue Combination. Now we can proceed to combining the contributions of different cues on the pixel level. For this, we adopt the idea of formulating hypothesis selection as a Quadratic Boolean Optimization Problem in an MDL framework [11]. Each hypothesis is evaluated in terms of the *savings* that can be obtained in the description of an image by explaining part of it by h. The savings of each hypothesis are expressed as

$$S_h = -\kappa_1 + (1 - \kappa_2) \frac{N}{A_\sigma} + \kappa_2 \frac{1}{A_\sigma} \sum_{\mathbf{p} \in Seg(h)} f(\mathbf{p}, h, Q) \quad (5)$$

where N is the number of pixels that can be explained by h, A_σ is its *expected area* at scale σ, κ_2 is a weighting factor to balance out the influence of a hypothesis's area versus its support in the image (left at a fixed value in our experiments), and κ_1 is the parameter over which the final performance curves are plotted. If multiple hypotheses overlap, their respective savings terms interact, since each pixel can only be assigned to a single hypothesis.

Depending on the definition of f, we can achieve different effects. The canonical way of combining the different cues would be to simply ignore possible correlations and marginalize over the cues q_m. This can be expressed by the following *sum* criterion:

$$f_{\text{sum}}(\mathbf{p}, h, Q) = \sum_m p(\mathbf{p} = figure | h, q_m) p(q_m). \quad (6)$$

However, this marginalization has the problem that it may reinforce local misclassifications if the cues are correlated. An opposite strategy is to completely remove correlation by only trusting the strongest cue. This leads to the *max* criterion:

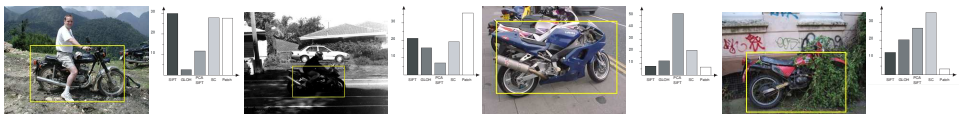


Figure 2: *Some detections and the corresponding relative cue contributions.*

$$f_{\max}(\mathbf{p},h,Q) = \max_m p(\mathbf{p} = \mathit{figure}|h,q_m)p(q_m). \tag{7}$$

However, this criterion is also problematic, since it relies on the assumption that all cues are well-behaved. If one or more cues respond too strongly to background structures, the whole system may become biased and additional false positives may be generated. For this reason, we also propose a third criterion, which is a combination of the two extremes. It builds the per-pixel *average* over all cues that are sufficiently confident, i.e. where $p(\mathbf{p} = \mathit{figure}|h,q_m) \gg p(\mathbf{p} = \mathit{ground}|h,q_m)$.

$$f_{\text{avg}}(\mathbf{p},h,Q) = \operatorname*{avg}_m p(\mathbf{p} = \mathit{figure}|h,q_m)p(q_m) \tag{8}$$

These criteria implement a highly flexible combination strategy. Instead of weighting each cue just by a fixed prior, they can decide for each image pixel anew which cues to consider, where the decision is made based on the cues' own confidence estimates. At the same time, eqs. (6) and (8) avoid putting all trust into a single cue that might bias the results negatively. Figure 1 summarizes the final cue combination procedure. The system first generates a set of hypotheses (Fig. 1(a)) by pooling the information from all cues. For each hypothesis, it then computes a top-down segmentation per cue (Fig. 1(b)), whereupon the verification criterion from eq. (3) is executed in order to fuse the individual cues' $p(\mathit{figure})$ probability maps (Fig. 1(c)) into a common system response.

Discussion and Analysis. It is important to emphasize the difference of the proposed cue integration scheme to the far simpler approach of running several region detectors in parallel and pooling their features in a common codebook (as used e.g. in [4]). If only a single kind of region descriptors is used, such an approach would be similar to our integration using the *sum* criterion. However, as soon as several different region descriptors shall be employed, a combination into a common codebook is no longer possible, since the different descriptors are not comparable. Our proposed approach, on the other hand, readily scales to this case and allows to combine the different cue contributions on a flexible per-pixel basis, which is something no other current approach can achieve.

The proposed cue integration scheme was motivated by the potential of different local cues to complement each other by interpreting the image information in different ways. In order to visualize that this can positively affect recognition performance, we introduce the following *argmax* criterion as an analysis tool.

$$f_{\text{argmax}}(\mathbf{p},h,Q) = \operatorname*{argmax}_m p(\mathbf{p} = \mathit{figure}|h,q_m)p(q_m) \tag{9}$$

This criterion selects for each hypothesis pixel the index of the most confident cue. Fig. 1(d) shows the resulting maps for the two example images, where each shade of gray corresponds to one of the five descriptors *SIFT*, *GLOH*, *PCA-SIFT*, *Shape Context*, and *Patch* (c.f. Sec. 4). These images are readily interpretable. For instance, it becomes evident that in the top example, the outer rim of the front wheel is best captured by *Shape Context* descriptors, while the wheel's hub is better represented by *GLOH*. In the bottom example, on the other hand, changed contrast to the background has modified the image content sufficiently, such that similar structures on the rear wheel are better captured by *SIFT*.

We can further quantify the relative importance of each cue to a particular hypothesis h by building up a histogram of their individual contributions. Fig. 1(e) shows the corresponding

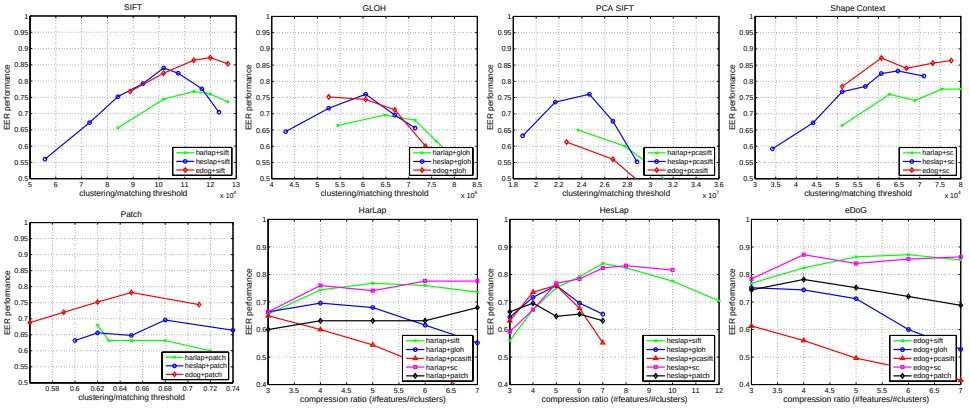


Figure 3: *Single-cue EER performances for all detector/descriptor combinations on the TUD motorbikes. The plots show the performance gradation when the clustering/matching threshold is varied. In all following experiments, we use only the best-performing parameter setting for each cue.*

cue importance histograms. As can be seen, the relative importance of the cues changes also quantitatively. Some more examples for different test images are shown in Fig. 2, further corroborating this observation.

4 Experimental Setup

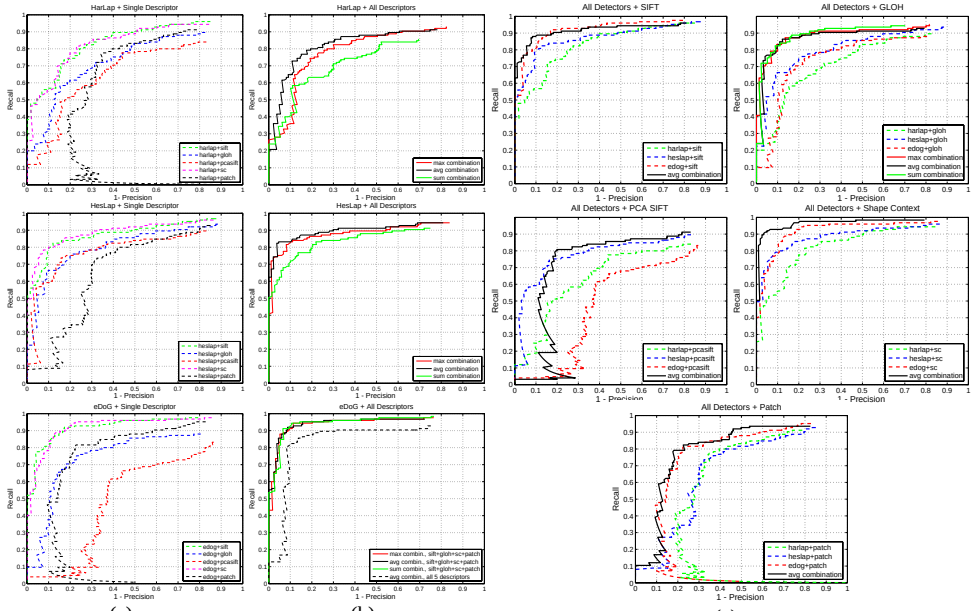
In the rest of the paper, we evaluate our proposed multi-cue integration method on real-world detection tasks. We first describe the selection of cues we build upon and the test data sets.

Interest Region Detectors. We compare three different scale-invariant interest region detectors. The *Harris-Laplace* and *Hessian-Laplace* detectors look for scale-adapted maxima of the Harris function and Hessian determinant, respectively [14], where the locations along the scale dimension are found by the Laplacian-of-Gaussian. The *DoG* detector [12] finds regions at 3D scale-space extrema of the Difference-of-Gaussian.

Region Descriptors. In addition, we evaluate five different region descriptors. *SIFT* descriptors [12] are 3D histograms of gradient locations and orientations with 4×4 location and 8 orientation bins. The resulting descriptor has 128 dimensions. *GLOH* descriptors [14] are an extension of *SIFT*. They use 17 location and 16 orientation bins organized in a log-polar grid. PCA is used to reduce the dimensionality to 128. *PCA-SIFT* [8] are vectors of image gradients in *x* and *y* direction sampled within the support region and reduced to 36 dimensions with PCA. *Shape Context* (*SC*) [2, 14] descriptors are histograms of gradient orientations sampled at edge points in a log-polar grid with 9 location and 4 orientation bins and thus 36 dimensions. For comparison, we include 25×25 pixel *Patches* [1, 10], which lead to a descriptor of length 625. This set of descriptors was explicitly chosen to sample different sources of information. *SIFT*, *GLOH*, and *PCA-SIFT* are based on gradient information; *SC* descriptors are based on edges; and *Patches* take the full image region into account.

The evaluation is performed with an own implementation of the *DoG* detector (denoted *eDoG* in the figures) and *Patch* descriptor. For all other detectors and descriptors, we used the implementations publicly available at [16]. Patches were compared using *Normalized Correlation*; all other descriptors were compared using Euclidean distances.

Training and Test Data. We first evaluate the different stages of our approach on the TUD motorbike set, which is part of the PASCAL collection [5]. This data set consists of 115 im-



(a) (b) (c)

Figure 4: *Cue combination performances on the TUD motorbikes: (a) single-cue performance; (b) performance of the different combination strategies using all 5 descriptors with the same detector; (c) cue combination performance when the same descriptors are applied to different detectors.*

ages containing a total of 125 motorbikes at different scales and with clutter and occlusion. Training is done on 153 motorbike side views from the CalTech training set [6] which are shown in front of uniform backgrounds allowing for easy segmentation. We then show that the results generalize also to other scenarios by applying the approach to two more challenging data sets using the same parameter settings. The first is the VOC motorbikes `test2` set, which has been used as a localization benchmark in the 2005 PASCAL Challenge [5]. This data set consists of 202 images containing a total of 227 motorbikes at different scales and seen from different viewpoints. Only about 37% of those motorbikes are shown in side views, though, thus limiting the maximally achievable recall for our system. Finally, we apply our method to the pedestrian test set from [11]. It consists of 209 images containing crowded scenes with a total of 595 pedestrians, mostly shown in side views but with significant overlap and occlusion. Training for this test is done on 216 side views of pedestrians for which a segmentation mask was available, using the same parameter settings as for the motorbike experiments. In all three cases, the task is to detect and localize the objects in the test images and determine their correct bounding boxes (using the evaluation criterion from [11] for the first and third test set, and the criterion from [5] for the second test set).

5 Results

Single-Cue Performance. In order to obtain an unbiased estimate of the cues' potentials, it is important to ensure that they are evaluated at their optimal setting. As a first step, we therefore evaluate each cue separately and try to find its performance optimum.

In our formulation of the approach, there is one open parameter that has to be adjusted for each cue, namely the question how much the clustering step should compress the training

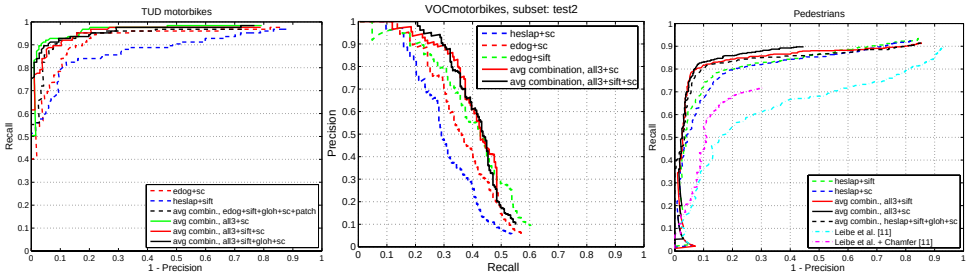


Figure 5: *Performance comparison on the TUD motorbikes (left), the more difficult VOC motorbike* `test2` *set (middle), and the pedestrian test set (right). The middle plot is rotated* 90° *to make it consistent with the ones in [5]. Please note that while our detector is exclusively trained on side views, only 39% of the motorbikes in the VOC set are shown in side views, thus limiting the maximally achievable recall.*

features during codebook generation. When using agglomerative clustering, this translates to the question how compact the codebook clusters should be for optimal performance. One option is to define a *minimum similarity* after which clustering should be stopped. Another option is to fix a certain *cluster compression ratio* (*#features*/*#clusters*). Previous evaluations [13] have favored the latter option, but it is not guaranteed that this choice is optimal.

In order to analyze the clustering/matching threshold's influence on recognition performance, we applied all 15 detector/descriptor combinations to the TUD motorbikes set and compared their equal error rate (EER) detection performance for 5–7 different threshold settings. Figure 3 shows the results of this experiment, both separated per descriptor and per detector. We can make two observations. First, when comparing descriptors across different detectors, a clear performance optimum can be found at a certain similarity for *SIFT*, *GLOH*, *PCA-SIFT*, and *SC*. The cluster compression ratio, on the other hand, does not seem to have a consistent influence. We can therefore formulate the recommendation to use the cluster similarity as a criterion for selecting the clustering level for those descriptors. Second, the results allow to rank the detector/descriptor combinations based on their single-cue performance. For the descriptors, *SIFT* and *SC* perform consistently best over all three detectors. For the detectors, *Hessian-Laplace* and *DoG* perform best in all but one case. In terms of combinations, *DoG+SIFT* and *DoG+SC* obtain the best performance with 87% EER.

Combining Different Descriptors. Next, we examine cue combination in a maximally correlated setting. For this, we apply all five region descriptors to the output of the same interest point detector and compare the performance of the three proposed combination strategies. The results of this experiment can be seen in Fig. 4(a,b). For *Harris-Laplace* and *Hessian-Laplace*, there is a significant difference between the three performance curves, with *sum* combination performing worst, then *max* combination, and *average* combination performing best. This confirms our expectations from Section 3. Compared to the best single-cue performance with *SIFT* or *SC* descriptors, *average* combination achieves a small performance increase from 77.6% to 80.0% (*Harris-Laplace*) and from 82.4% to 85.6% EER (*Hessian-Laplace*), respectively. For *DoG*, a significant performance increase from 87.2% to 91.2% EER can be shown if all descriptors except *PCA-SIFT* are combined. Including *PCA-SIFT* degrades overall performance to 85.6%, suggesting that those descriptors are not as informative as the others, perhaps because of their projection onto a general-purpose PCA basis.

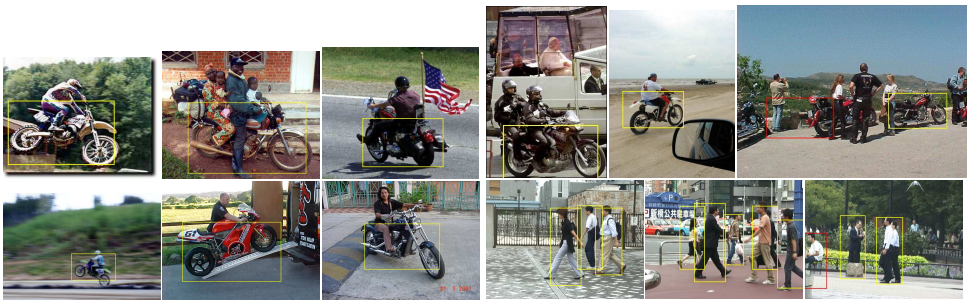


Figure 6: *Example multi-cue detections of our approach on difficult images from the VOC motorbikes and the pedestrian set (at the EER).*

Combining Different Detectors. The opposite experiment is to apply the same descriptors to three different region detectors and compare the combined performances. This is shown in Fig. 4(c). As there are only small differences between the performance of the three combination strategies, we just display the curve for *average* combination in order to reduce clutter. The most remarkable observation from this experiment is the improvement of over 10% EER obtained by the *GLOH* descriptors from 76.0% to 86.4%. Apparently, this descriptor benefits most from additional samples in the image. In contrast, *SIFT* shows only a small improvement to 88.8% EER. The best absolute performance is achieved by the *SC* combination with 92.8% EER. The *PCA-SIFT* and *Patch* descriptors, finally, do not profit from the evaluated combination.

Full Multi-Cue Combination. Finally, we present results combining multiple detectors and multiple descriptors at the same time. Fig. 5(left) compares the performance of *SIFT+SC* and *SIFT+GLOH+SC* with all three detectors. Although those combinations do not increase EER performance any more, further improvement can be observed in terms of precision. In particular, recall at the zero-false-positive level is increased from 50% (only *SC*) over 62% (*SIFT+SC*) to 75% (all three descriptors). This is an important result, since high precision is a prerequisite for many real-world applications.

In order to ensure that the results generalize also to different settings, we apply our multi-cue approach to the more challenging VOC motorbikes set using the same parameter settings as for the first experiments. Fig. 5(middle) shows the results of this experiment. As can be seen from the plot, the combination of multiple cues again improves performance and increases the detection precision considerably. As a comparison with [5] shows, it is the best result reported for this data set so far. The best combination of *SIFT+SC* achieves 21% recall with zero false positives and scales up to 30% recall at 90% precision. Considering that the test set contains only about 39% side views, this is an excellent result. Fig. 6 visualizes the range of motorbike appearances that are still reliably detected by our approach. Although the system has only been trained on a single viewpoint, the increased robustness from multi-cue integration makes it possible to compensate for a certain level of out-of-plane rotation.

Last but not least, we apply our multi-cue approach to the pedestrian test set from [11] using the same clustering/matching thresholds as for the motorbikes. The results are shown in Fig. 5(right). Again, the combination of multiple cues increases performance significantly from 80% EER for the best single cues to 84.7% for *SC* with all three detectors and to 82.6% with *HesLap* with *SIFT+GLOH+SC*. In comparison, we show the results from [11], which are clearly outperformed by our multi-cue system.

6 Discussion & Conclusion

In conclusion, we have proposed a robust and flexible multi-cue integration scheme that operates even when the cues are highly correlated. It has been shown to improve performance consistently on three different data sets and for two different categories. The improvement is particularly visible in terms of recognition precision and, for the motorbike test sets, high recall values at the zero-false-positive level. Compared to a canonical cue combination strategy of simply adding the weighted cue responses, our proposed approach can react more flexibly to varying cue performance and adapt itself automatically. This advantage could also be verified quantitatively in cases where the cues were strongly correlated.

In order to further evaluate its performance we have conducted an extensive study, comparing 3 state-of-the-art interest region detectors and 5 different descriptors in the context of multi-cue integration. The results of this evaluation allow to rank the cues both based on their individual performance and their suitability for integration. In addition, we can draw several interesting conclusions. When set to the right clustering level, *SIFT* and *SC* features performed consistently better than all other descriptors in this evaluation. In addition, feature combinations with either *SC* descriptors and several different region detectors or *DoG*/*Hessian*-*Laplace* regions with several different descriptors achieved the highest overal performance level. These two extremes thus provide an axis along which the set of cues can be varied depending on implementation tradeoffs (i.e. either sampling more points or using the sampled information more efficiently).

Acknowledgments. This work has been funded, in part, by the EU projects COSY (IST-2002-004250) and DIRAC (IST-2005-27787).

References

[1] S. Agarwal, A. Atwan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11):1475–1490, 2004.
[2] S. Belongie, J. Malik, and J. Puchiza. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, April 2002.
[3] C. Brautigam, J.-O. Eklund, and H. Christensen. A model-free approach for integrating multiple cues. In *ECCV'98*, 1998.
[4] G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *ICCV'03*, 2003.
[5] M. Everingham et al. (34 authors). The 2005 pascal visual object class challenge. In *Selected Proceedings of the 1st PASCAL Challenges Workshop*, LNAI. Springer, to appear. `http://www.pascal-network.org/challenges/VOC/`.
[6] R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *CVPR'03*, 2003.
[7] A. Garg, S. Agarwal, and T. Huang. Fusion of global and local information for object detection. In *ICPR'02*, 2002.
[8] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *CVPR'04*, 2004.
[9] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.
[10] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Workshop on Stat. Learn. in Comp. Vis.*, 2004.
[11] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR'05*, 2005.
[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
[13] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *ICCV'05*, 2005.
[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):31–37, 2005.
[15] M.E. Nilsback and B. Caputo. Cue integration through discriminative accumulation. In *CVPR'04*, 2004.
[16] Oxford interest point webpage. `http://www.robots.ox.ac.uk/~vgg/research/affine/`.
[17] Z. Sun. Adaptation for multiple cue integration. In *CVPR'03*, 2003.
[18] J. Triesch and C. Eckes. Object recognition with multiple feature types. In *ICANN'98*, 1998.
[19] J. Triesch and C. von der Malsburg. Democratic integration: Self-organized integration of adaptive cues. In *Neural Computation*, pages 2049–2074, 2001.
[20] K. Woods, W.P. Kegelmeyer Jr., and K. Bowyer. Combination of multiple classifiers using local accuracy estimation. *PAMI*, 19(4):405–410, 1997.