Building Detection from Mobile Imagery Using Informative SIFT Descriptors*

Gerald Fritz, Christin Seifert, Manish Kumar, and Lucas Paletta

JOANNEUM RESEARCH Forschungsgesellschaft mbH,
Institute of Digital Image Processing,
Wastiangasse 6, A-8010 Graz, Austria
lucas.paletta@joanneum.at

Abstract. We propose reliable outdoor object detection on mobile phone imagery from off-the-shelf devices. With the goal to provide both robust object detection and reduction of computational complexity for situated interpretation of urban imagery, we propose to apply the 'Informative Descriptor Approach' on SIFT features (i-SIFT descriptors). We learn an attentive matching of i-SIFT keypoints, resulting in a significant improvement of state-of-the-art SIFT descriptor based keypoint matching. In the off-line learning stage, firstly, standard SIFT responses are evaluated using an information theoretic quality criterion with respect to object semantics, rejecting features with insufficient conditional entropy measure, producing both sparse and discriminative object representations. Secondly, we learn a decision tree from the training data set that maps SIFT descriptors to entropy values. The key advantages of informative SIFT (i-SIFT) to standard SIFT encoding are argued from observations on performance complexity, and demonstrated in a typical outdoor mobile vision experiment on the MPG-20 reference database.

1 Introduction

Research on visual object detection has recently focused on the development of local interest operators [7, 9, 11, 6, 5] and the integration of local information into robust object recognition [1, 6]. Recognition from local information serves several purposes, such as, improved tolerance to occlusion effects, or to provide initial evidence on object hypotheses in terms of providing starting points in cascaded object detection. Recently, [11] investigated informative image fragments for object representation and recognition, and [3] applied information theoretic analysis to determine saliency measures in multi-view object recognition. While these approaches performed fundamental analyses on the appearance patterns, the natural extension of improving advanced local detectors by investigating

* This work is supported by the European Commission funded projects MACS under grant number FP6-004381 and MOBVIS under grant number FP6-511051, and by the FWF Austrian Joint research Project Cognitive Vision under sub-projects S9103-N04 and S9104-N04.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 629–638, 2005.

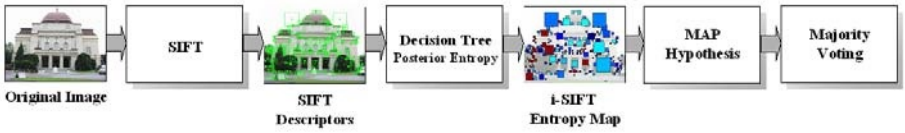


Fig. 1. Concept for automated building recognition. First, standard SIFT descriptors are extracted within the test image. The proposed informative SIFT (i-SIFT) approach determines the entropy in the descriptor and performs decision making (MAP hypothesizing) only on attended descriptors. Majority voting is then used to integrate local votes into a global classification

about the information content they provide with respect to object discrimination remained open. The Informative Feature Approach is particularly suited for computer vision on emerging technologies, such as, mobile devices, requiring careful outline of algorithms to cope with limited resources, crucial constraints on response times, and complexity in the visual input from real world conditions.

The key contribution of the presented work is (i) to demonstrate a reliable methodology for the application of object detection in mobile phone imagery, and (ii) illustrating that the Informative Feature Approach [3] can perfectly be extended to complex features, such as the SIFT interest point detector [6], to render recognition more efficient. First, we provide a thorough analysis on the discriminative power of complex SIFT features, using local density estimations to determine conditional entropy measures, that makes the actual local information content explicit for further processing. Second, we build up an efficient i-SIFT based representation, using an information theoretic saliency measure to construct a sparse SIFT descriptor based object model. Rapid SIFT based object detection is then exclusively applied to test patterns with associated low entropy, applying an attention filter with a decision tree encoded entropy criterion. We demonstrate that i-SIFT (i) provides better foreground-background discrimination, (ii) significantly reduces the descriptor dimensionality, (iii) decreases the size of object representation by one order of magnitude, and (iv) performs matching exclusively on attended descriptors, rejecting the majority of irrelevant descriptors.

The experiments were performed on raw mobile phone imagery on urban tourist sights under varying environment conditions (changes in scale, viewpoint, and illumination, severe degrees of partial occlusion). We demonstrate in this challenging outdoor object detection task the superiority in using informative SIFT (i-SIFT) features to standard SIFT, by increased reliablity in foreground/background separation, and the significant speedup of the algorithm by one order of magnitude, requiring a fraction ($\approx 20\%$) of features of lower dimensionality (30%) for representation.

2 Informative Local Descriptors

The *Informative Descriptor Approach* requires to extract relevant features in a pre-processing stage to recognition, by optimizing feature selection with respect

to the information content in the context of a specific task, e.g., object recognition. In the following sections, we motivate the determination of informative descriptors from information theory (Sec. 2.1), and describe the application to local SIFT descriptors (Sec. 2.2).

2.1 Local Information Content

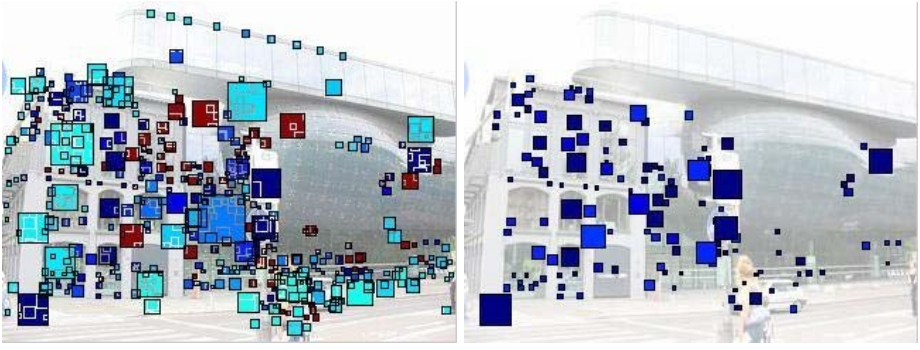
Saliency of interest points has been attributed due to various interpretations, such as, from sample densities within local neighborhoods [4], or according to the class specific general frequency of occurrence in recognition [11]. The *Informative Feature Approach* determines the information content from a posterior distribution with respect to given task specific hypotheses. In contrast to costly *global* optimization, we expect that it is sufficiently accurate to estimate a *local* information content, by computing it from the posterior distribution within a sample test point's local neighborhood in feature space.

We are primarily interested to get the *information content* of any sample local descriptor $\mathbf{f}_i$ in feature space $\mathcal{F}$, $\mathbf{f}_i \in \mathcal{R}^{|\mathcal{F}|}$, with respect to the task of object recognition, where o_i denotes an object hypothesis from a given object set Ω. For this, we need to estimate the entropy $H(O|\mathbf{f}_i)$ of the posterior distribution $P(o_k|\mathbf{f}_i)$, $k = 1 \ldots \Omega$, Ω is the number of instantiations of the object class variable O. The Shannon conditional entropy denotes $H(O|\mathbf{f}_i) \equiv -\sum_k P(o_k|\mathbf{f}_i) \log P(o_k|\mathbf{f}_i)$. We approximate the posteriors at $\mathbf{f}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window of a local neighborhood ϵ, $||\mathbf{f}_i - \mathbf{f}_j|| \leq \epsilon, j = 1 \ldots J$. We weight the contributions of specific samples $\mathbf{f}_{j,k}$ - labeled by object o_k - that should increase the posterior estimate $P(o_k|\mathbf{f}_i)$ by a Gaussian kernel function value $\mathcal{N}(\mu, \sigma)$ in order to favor samples with smaller distance to observation $\mathbf{f}_i$, with $\mu = \mathbf{f}_i$ and $\sigma = \epsilon/2$. The estimate about the Shannon conditional entropy $\hat{H}(O|\mathbf{f}_i)$ provides then a measure of ambiguity in terms of characterizing the information content with respect to object identification within a single local observation $\mathbf{f}_i$. Fig. 2 depicts *discriminative descriptors* in an entropy-coded representation of local SIFT features $\mathbf{f}_i$.

From discriminative descriptors we proceed to *entropy thresholded* object *representations*, providing increasingly sparse representations with increasing recognition accuracy, in terms of storing only *selected* descriptor information that is *relevant for classification* purposes, i.e., those $\mathbf{f}_i$ with $\hat{H}(O|\mathbf{f}_i) \leq \Theta$. A specific choice on the threshold Θ consequently determines both storage requirements and recognition accuracy (Sec. 4). To speed up the matching we use efficient memory indexing of nearest neighbor candidates described by the adaptive K-d tree method [2].

2.2 Informative SIFT Descriptors

We apply the *Informative Feature Approach* on SIFT based descriptors that are among the best local descriptors with respect to matching distinctiveness, invariance to blur, image rotation, and illumination changes [8]. However, critical bottlenecks in SIFT based recognition are identified as performing extensive



(a) (b)

Fig. 2. Informative descriptors for saliency

SIFT keypoint matching with high computational complexity due to the nearest neighbor indexing, and the lack of any representation of uncertainty to enable approximate reasoning. We apply informative feature selection to the SIFT descriptor with the aim to significantly decrease the computational load using attentive matching, while attaining improved detection accuracy, and providing a probabilistic framework for individual SIFT descriptors.

SIFT Descriptors. Descriptors of the Scale Invariant Feature Transform (SIFT [6]) are invariant to image scale and rotation, in addition, they show robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. [6] further augments the SIFT descriptor into a distinctive feature approach, by proposing specific matching and descriptor representation methods techniques for object recognition. While the informative SIFT (i-SIFT) approach will specifically improve the matching and representation procedures regarding the complete SIFT approach, any consecutive recognition methodology mentioned in [6] might be applied as well to i-SIFT, such as, using the Hough transform to identify clusters belonging to a single object, etc.

i-SIFT Descriptors. The application of the *Informative Descriptor Approach* tackles three key aspects of SIFT estimation: (i) reducing the high dimensionality (128 features) of the SIFT keypoint descriptor, (ii) thinning out the number of training keypoints using posterior entropy thresholding (Sec. 2.1), in order to obtain an informative and sparse object representation, and (iii) providing an entropy sensitive matching method to reject non-informative outliers, described in more detail as follows,

1. *Reduction of high feature dimensionality* (128 features) of the SIFT descriptor is crucial to keep nearest neighbor indexing computationally feasible. Possible solutions are K-d and Best-Bin-First search ([6]) that practically perform by $\mathcal{O}(ND)$, with N training prototypes composed of D features.

To discard statistically irrelevant feature dimensions, we applied Principal Component Analysis (PCA) on the SIFT descriptors. This is in contrast to the PCA-SIFT method [5], where PCA is applied to the normalized gradient pattern, but that also becomes more errorprone under illumination changes [8].

2. *Information theoretic selection of representation candidates.* According to the Informative Descriptor Approach (Sec. 2.1) we exclusively select *informative* local SIFT descriptors for object representation. The degree of reduction in the number of training descriptors is determined by threshold Θ for accepting sufficiently informative descriptors. In the experiments (Sec. 4) this approximately reduces the representation size by one order of magnitude. *Object sensitive informative descriptor selection* avoids a general cut-off determined by the posterior entropy measure but attributes to each objects its partition $1/|\Omega|N_{sel}$ of a predetermined total number N_{sel} of to-be selected descriptors. This prevents the method from associating too few SIFT descriptors to a corresponding object representation.
3. *Entropy sensitive matching* in nearest neighbor indexing is then necessary as a means to reject outliers in analyzing test images. Any test descriptor $\mathbf{f}_*$ will be rejected from matching if it comes not close enough to any training descriptor $\mathbf{f}_i$, i.e., if $\forall \mathbf{f}_i : |\mathbf{f}_i - \mathbf{f}_*| < \epsilon$, and ϵ was determined so as to optimize posterior distributions with respect to overall recognition accuracy (Sec. 2.1).

3 Attentive Object Detection

This section outlines a framework for object detection that enables performance comparison between SIFT and i-SIFT based local descriptors. i-SIFT based object detection can achieve a significant speedup from attentive filtering for the rejection of less promising candidate descriptors. This attentive mapping of low computational complexity is described in terms of a decision tree which learns its tree structure from examples, requiring very few attribute comparisons to decide upon acceptance or rejection of a SIFT descriptor under investigation.

Object Recognition and Detection. To enable direct performance comparison between SIFT and i-SIFT based object recognition, we determine an adapted majority voting procedure to decide upon a preferred object hypothesis. Detection tasks require the rejection of images whenever they do not contain any objects of interest. For this we consider to estimate the entropy in the posterior distribution - obtained from a normalized histogram of the object votes - and reject images with posterior entropies above a predefined threshold. The proposed recognition process is characterized by an entropy driven selection of image regions for classification, and a voting operation, as follows,

1. **Mapping** of local patterns into descriptor subspace.
2. **Probabilistic interpretation** to determine local information content and associated entropy measure.

3. **Rejection** of descriptors contributing to ambiguous information.
4. **Nearest neighbor analysis** of selected imagettes within ϵ-environment .
5. **Majority voting** for object identifications over a region of interest.

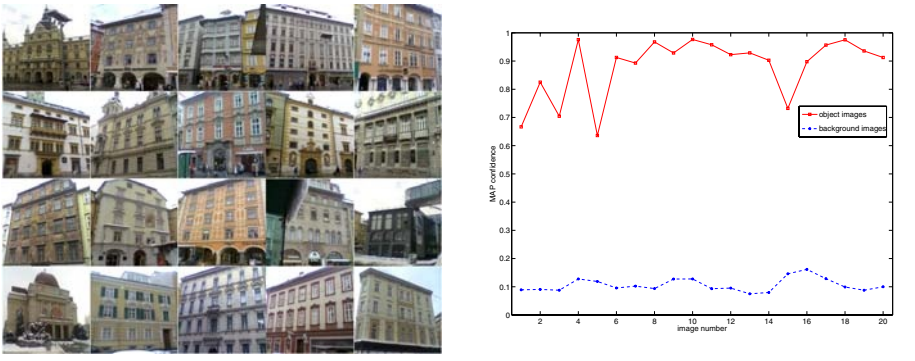
Each pattern from a test image that is mapped to SIFT descriptor features is analyzed for its conditional entropy with respect to the identification of objects $o_i \in O$. An entropy threshold Θ for rejecting ambiguous test descriptors in eigenspace is most easily identical with the corresponding threshold applied to get a sparse model of reference points. Object recognition on a collection of (matched and therefore labelled) SIFT descriptors is then performed on finding the object identity by majority voting on the complete set of class labels attained from individual descriptor interpretations.

Learning Attentive Matching. For a rapid estimation of local entropy quantities, the descriptor encoding is fed into the decision tree which maps SIFT descriptors $\mathbf{f}_i$ into entropy estimates $\hat{H}$, $\mathbf{f}_i \mapsto \hat{H}(\Omega|\mathbf{f}_i)$. The C4.5 algorithm [10] builds a decision tree using the standard top-down induction of decision trees approach, recursively partitioning the data into smaller subsets, based on the value of an attribute. At each step in the construction of the decision tree, C4.5 selects the attribute that maximizes the information gain ratio. The induced decision tree is pruned using pessimistic error estimation [10]. The extraction of informative SIFTs (i-SIFTS) in the image is performed in two stages. First, the decision tree based entropy estimator provides a rapid estimate of local information content of a SIFT key under investigation. Only descriptors $\mathbf{f}_i$ with an associated entropy below a predefined threshold $\hat{H}(O|\mathbf{f}_i) < \Theta$ are considered for recognition. Only these selected discriminative descriptors are then processed by nearest neighbor analysis, with respect to the object models, and interpreted via majority voting.

Computational Complexity. There are several issues in using i-SIFT attentive matching that significantly ease the resulting computational load, showing improvements along several dimensions. Firstly, information theoretic selection of candidates for object representation experimentally *reduces the size* of the object representation of up to *one order of magnitude*, thus supporting sparse representations on devices with limited resources, such as, mobile vision enhanced devices. Secondly, the reduction of dimensionality in the SIFT descriptor representation may in addition *decrease computational load down to* $\leq 30\%$. Finally, the attentive decision tree based mapping is applied to reject SIFT descriptors for further analysis, thereby *retaining only about* $\leq 20\%$ SIFT descriptors for further analysis.

4 Experiments

Targeting emerging technology applications using computer vision on mobile devices, we perform the performance tests on mobile phone imagery captured



(a) (b)

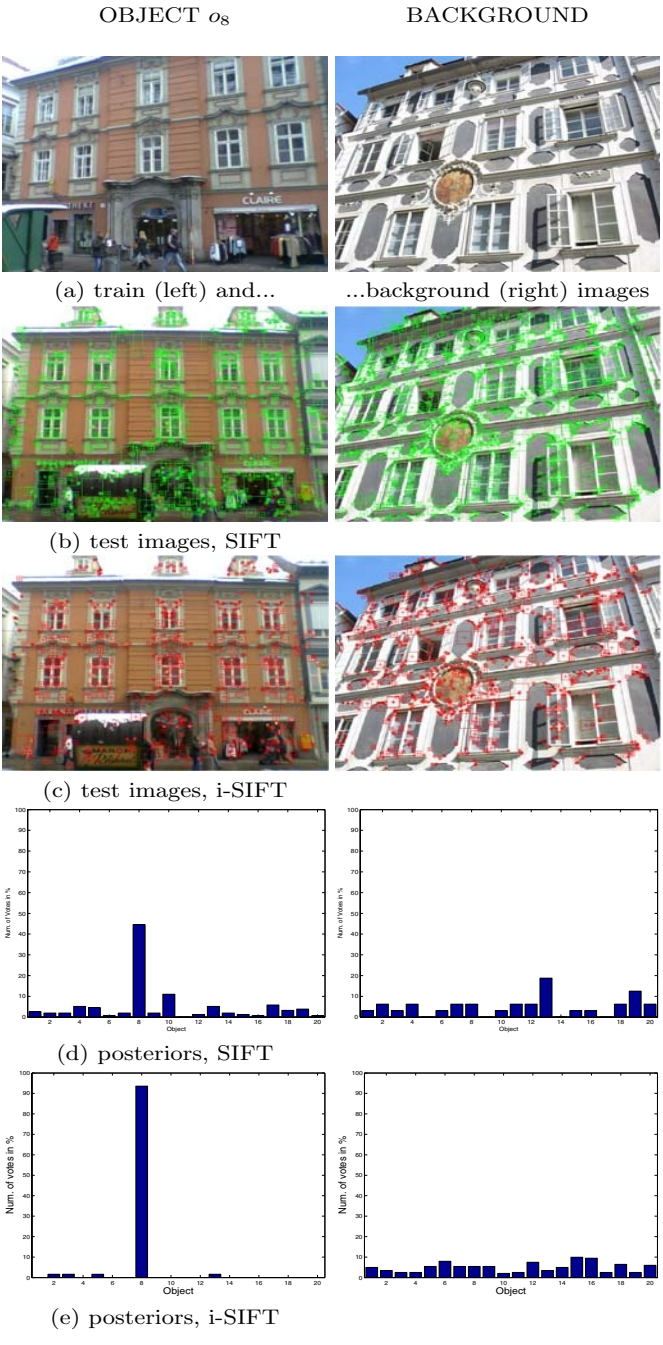
Fig. 3. The MPG-20 database, consisting of mobile phone images from 20 buildings (numbered o_1–o_{20} from top-left to bottom-right) in the city of Graz (displayed images were used for training, see Sec. 4). (b) i-SIFT outperforming SIFT supported MAP confidence based discrimination between object and background

about tourist sights in the urban environment of the city of Graz, Austria, i.e., from the MPG-20 database (see below, Fig. 3). In order to evaluate the improvements gained from the 'Informative Descriptor Approach', we compare the performance between the standard SIFT key matching and the i-SIFT attentive matching.

MPG-20 Database. The MPG-20 database[1] includes images from 20 objects, i.e., facades of buildings from the city of Graz, Austria. Most of these images contain a tourist sight, together with 'background' information from surrounding buildings, pedestrians, etc. The images were captured from an off-the-shelf camera phone (Nokia 6230) of resolution 640×480, containing severe changes in 3D viewpoint, partial occlusions, scale changes by varying distances for exposure, and various illumination changes due to different weather situations and changes in daytime. For each object, we then selected 2 images taken by a viewpoint change of $\approx \pm 30°$ of a similar distance to the object for training to determine the i-SIFT based object representation. 2 additional views - two different front views of distinct distance and therefore significant scale change - were taken for test purposes, giving 40 test images in total. Further images (noted in separation) were obtained from 'background', such as, other buildings, landscape, etc., and from objects under severe illumination conditions (in the evening, Christmas lighting, etc.).

Standard SIFT Based Key Matching. The training images were bit-masked by hand, such that SIFT descriptors on background information (cars, surrounding

[1] The MPG-20 (Mobile Phone imagery Graz) database can be downloaded at the URL http://dib.joanneum.at/cape/MPG-20.



(a) train (left) and... ...background (right) images

(b) test images, SIFT

(c) test images, i-SIFT

(d) posteriors, SIFT

(e) posteriors, i-SIFT

Fig. 4. Sample object detection results for object o_8 (left) and background (right), (a) depicting train images, (b) SIFT descriptor locations on test images, (c) selected i-SIFT descriptors, (d) posterior distribution on object hypotheses from SIFT and (e) i-SIFT descriptors, demonstrating more distinctive results for i-SIFT based interpretation

Table 1. Performance comparison between *standard SIFT* keypoint matching [6] and *i-SIFT* attentive matching on MPG-20 mobile imagery

Recognition Method	MAP accuracy MPG-20 [%]	PT [%]	PF [%]	obj $\bar{H}$	bgd $\bar{H}$	obj avg. MAP	bgd avg. MAP
SIFT	95.0	82.5	0.1	3.0	3.4	43.9	18.7
i-SIFT	97.5	100.0	0.0	0.5	4.1	88.0	10.6

buildings, pedestrians) were discarded. In total 28873 SIFT descriptors were determined for the 40 training images, 722 on average. The 40 (non-masked) test images generated a similar number of SIFT descriptors per image. For each of these descriptors the distances to the closest nearest neighbor (d_1) and the 2nd closest neighbor (d_2) was calculated. If the distance ratio ($\frac{d_1}{d_2}$) was greater than 0.8 the sift descriptor remained unlabeled (as described in [6], otherwise the label of the closest nearest neighbor was assigned. Thus, on average $\approx 30\%$ of the SIFT features were retained for voting. Object recognition is then performed using majority voting. The average entropy in the posterior of the normalized voting histograms was $H_{avg} \approx 3.0$. A threshold of 25% in the MAP hypothesis confidence was used as decision criterion to discriminate between object ($> 25\%$) and background ($\leq 25\%$) images (for both SIFT and i-SIFT).

i-SIFT Attentive Key Matching. For the training of the i-SIFT selection, the SIFT descriptor was projected to an eigenspace of dimension 40, thereby decreasing the original descriptor input dimensionality (128 features) by a factor of three. The size ϵ of the Parzen window for local posterior estimates was chosen 0.4, and 175 SIFT keys per object were retained for object representation. The threshold on the entropy criterion for attentive matching was defined by $\Theta = 1.0$. In total, the number of attended SIFT descriptors was 3500, i.e., $\approx 12.1\%$ of the total number that had to be processed by standard SIFT matching. The recognition accuracy according to MAP (Maximum A Posteriori) classification was 100%, the average entropy in the posterior distribution was $H_{avg} \approx 0.5$, in very analogy to the value achieved by standard SIFT matching (see above).

Table 1 illustrates the results of the MPG-20 experiments, and the results of a comparison between standard SIFT keypoint matching and i-SIFT attentive matching.

5 Summary and Conclusions

The presented work proposed a methodology for reliable urban object detection from off-the-shelf mobile phone imagery. We applied the *Informative Descriptor Approach* significantly improving the efficiency in object detection, both with respect to memory resources and to speedup the recognition process. The paper also introduces *attentive matching* using the informative SIFT (i-SIFT) descriptors, applying an information theoretic criterion for the selection of discrim-

inative SIFT descriptors. Matching with the i-SIFT descriptor (i) significantly reduces the dimensionality of the descriptor encoding, (ii) provides sparse object representations that reduce storage by one order of magnitude with respect to standard SIFT, and (iii) enables attentive matching by requiring $4-8$ times less SIFT features per image to be identifed by more costly nearest neighbor search.

This innovative local descriptor is most appropriate for sensitive operation under limited resources, such as, in mobile devices. We evaluated the performance of the i-SIFT descriptor on the public available MPG-20 database, including images from 20 building objects and 'non-object background' pictures from the city of Graz. The i-SIFT did not only compare well with the high recognition accuracy when using standard keypoint matching, but also provided discriminative posterior distributions, robust background detection, and - as surplus - significant speedup in processing times.

References

1. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2003.
2. J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226, 1977.
3. G. Fritz, L. Paletta, and H. Bischof. Object recognition using local information content. In *Proc. International Conference on Pattern Recognition, ICPR 2004*, volume II, pages 15–18. Cambridge, UK, 2004.
4. D. Hall, B. Leibe, and B. Schiele. Saliency of interest points under scale changes. In *Proc. British Machine Vision Conference, BMVC 2002*, Cardiff, UK, 2002.
5. Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proc. Computer Vision and Pattern Recognition, CVPR 2004*, volume 2, pages 506–513, Washington, DC, 2004.
6. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
7. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. European Conference on Computer Vision*, pages 128–142, 2002.
8. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. http://www.robots.ox.ac.uk/ vgg/research/affine/, 2004.
9. S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *Proc. British Machine Vision Conference*, pages 113–122, 2002.
10. J.R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
11. M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classication. In *Proc. International Conference on Computer Vision, ICCV 2003*, pages 281–288. Nice, France, 2003.