A Mobile Vision Service for Multimedia Tourist Applications in Urban Environments

Paletta L., Fritz G., Seifert C., Luley P., and Almer A.

***Abstract*— We present a computer vision system for the detection and identification of urban objects from mobile phone imagery, e.g., for the application of tourist information services. Recognition is based on MAP decision making over weak object hypotheses from local descriptor responses in the mobile imagery. We present an improvement over the standard SIFT key detector [1] by selecting only informative (i-SIFT) keys for descriptor matching. Selection is applied first to reduce the complexity of the object model and second to accelerate detection by selective filtering. We present results on the MPG-20 mobile phone imagery with severe illumination, scale and viewpoint changes in the images, performing with $\approx 98\%$ accuracy in identification, efficient (100%) background rejection, efficient (0%) false alarm rate, and reliable quality of service under extreme illumination conditions, significantly improving standard SIFT based recognition in every sense, providing – important for mobile vision – runtimes which are ≈ 8 (≈ 24) times faster for the MPG-20 (ZuBuD) database.**

I. Introduction

With the industrial miniaturization of cameras and mobile devices, the generation of and access to digital visual information has become ubiquitous. Today, most cameras are sold within mobile phones, accompanying the nomadic pedestrian through everyday life, in particular, in its urban environment. Computer vision could play a key role in using billions of images as a cue for context and object awareness, positioning, inspection, and annotation in general.

The original contribution of this paper is to provide a generic technology for the recognition of urban objects, i.e., buildings, in terms of a reliable mobile vision service in tourist information systems. A mobile user directing its mobile camera to an object of interest (Fig. 1) will receive annotation about location relevance (e.g., tourist sight) and the identity of the building, enabling the user to access choices on more detailed, object specific information.

Urban recognition has been approached with respect to categorical detection of architecture from line based features proposed by [2]. [3] presented a framework for structure recovery that aims at the same time towards posterior building recognition. [4] provided the first innovative attempt on building identification proposing local affine features for object matching. [5] introduced image retrieval methodology for the indexing of visually relevant information from the web for mobile location recognition. Following these merely conceptual approaches we propose an accurately and reliably working recognition service, providing detailed information on performance evaluation, both on mobile phone imagery and a reference building image database (Sec. V).

Our detection system grounds recognition on a MAP decision making on weak object hypotheses from local descriptor responses in the mobile imagery. We present an improvement over the standard SIFT key detector [1] by selecting only informative (i-SIFTs) keys for descriptor matching (Sec. III). Selection is applied first to reduce the complexity of the object model and second to accelerate detection by selective attention. We trained a decision tree to rapidly and efficiently estimate a SIFT's posterior entropy value, retaining only those keys for thorough analysis and voting with high information content (Sec. IV).

The experiments were performed on typical, low quality mobile phone imagery on urban tourist sights under varying environment conditions (changes in scale, viewpoint, illumination, varying degrees of partial occlusion). We demonstrate in this challenging outdoor object detection task the superiority in using informative SIFT (i-SIFT) keys to standard SIFT using the MPG-20 mobile phone image database, reporting increased reliability in object/background separation, accurate object identification, and providing a confidence quality measure that enables a highly stable mobile vision service.

II. Mobile Computer Vision System

Image based recognition provides the technology for both object awareness and positioning. Outdoor geo-referencing still mainly relies on satellite based signals where problems arise when the user enters *urban canyons* and the availability of satellite signals dramatically decreases due to various shadowing effects [6]. Alternative concepts for localization are economically not affordable, such as, INS and markers that need to be massively distributed across the urban area.

Fig. 1 depicts the technical concept and the three major stages in situated mobile object recognition and annotation. The system consists of an off-the-shelf camera-equipped smartphone, a GPS device (built-in, e.g., A-GPS, or Bluetooth externally connected), and a server accessible trough mobile services that runs the object recognition and annotation software. This specific client-server architecture enables large-scale application of urban object awareness, using GPS to index into the geo-referenced object database, and leaving object recognition restricted to a local urban area on the server. In the future, mobile clients might run the application even faster.

This work is funded by the European Commission's projects *MOBVIS* (FP6-511051) and *MACS* (FP6-004381), and by the FWF Austrian joint research project *Cognitive Vision* under sub-projects S9103-N04 and S9104-N04.

All authors are with the JOANNEUM RESEARCH Forschungsgesellschaft mbH, Institute of Digital Image Processing, Graz, A-8010, Austria. Corresponding author: `lucas.paletta@joanneum.at`

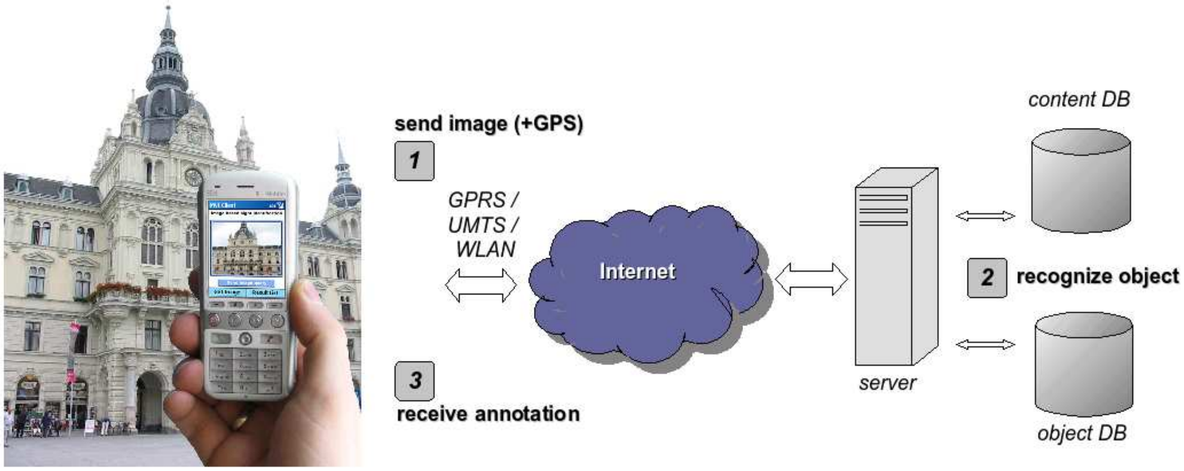


Fig. 1. Client-server architecture for object awareness in urban environments. (1) Images from the mobile devices are transferred to the server for (2) recognition and content information. (3) User receives annotation about object and locations.

Mobile recognition system In the first stage (Fig. 1-1), the user captures an image about an object of interest in its field of view, and a software client initiates submission of the image to the server. The transfer of the visual information to the server is performed either via GPRS, UMTS, WLAN (PDAs), or MMS (multimedia messaging service). If a GPS device (bluetooth or built-in A-GPS) is available, the smartphone reads the actual position estimate together with a corresponding uncertainty measure, and sends this together with the image to the server. In the second stage (Fig. 1-2), the web-service reads the message and analyzes the geo-referenced image. Based on a current quality of service and the given decision for object detection and identification, the server prepares the associated annotation information from the content database and sends it back to the client for visualization (Fig. 1-3).

Geo-contextual cueing *Global* object search in urban environments – comprising thousands of buildings – is a challenging research issue. However, within most application scenarios, positions would be available from GPS based geo-referencing, which can be used to index into an otherwise huge set of object hypotheses. Geo-reference indexing for selected object hypotheses first requires a database containing on-site captured geo-referenced imagery about objects. 'Ground truth' geo-referencing can be performed manually , e.g., on corresponding air-borne imagery (Fig. 2). From the differences between 'true' and on-site measured positions we can determine the average positioning error $\bar{\eta}$. Based on this quantity, we partition the complete set of object hypotheses into subsets of hypotheses ('neighborhood cell') of local context within a neighborhood $\bar{\eta} + \delta$ for further processing. For each of these neighborhoods, we would learn the informative features and an attentive mapping to saliency (Sec. III).

Situated object recognition In recognition mode, the GPS signal receiver firstly returns an on-site position estimate. We add then the radial distance of the uncertainty estimate $\bar{\varepsilon}$ to receive the urban area that most probably will contain the object under investigation (Fig. 2). We index now into the geo-referenced object database and receive the corresponding 'neighborhood cell' from which we derive the set of object hypotheses for accurate object identification.

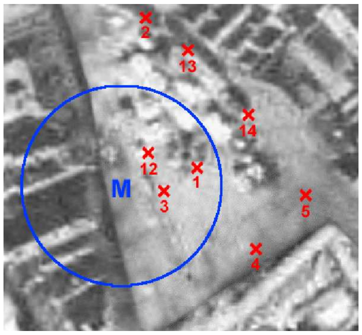


Fig. 2. Object hypothesis selection using geo-contextual information (overlaid on airborne image) from GPS based measurements (M). Positions of image acquisition of objects in the MPG-20 database (crosses) are indexed using the radial distance of the mean geo-referencing error.

The methodology is described as follows, Sec. III will present object representations by informative descriptors, Sec. IV describes attentive detection and identification, and Sec. V presents experimental results on mobile imagery with varying viewpoints and illumination conditions.

III. Informative Local Descriptors

Research on visual object detection has recently focused on the development of local interest operators [7], [8], [9], [1] and the integration of local information into robust object recognition [10], [9], [1]. 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. The SIFT (Scale Invariant Feature Transformation) descriptor [1] is widely used for its capabilities for robust matching to the recordings in

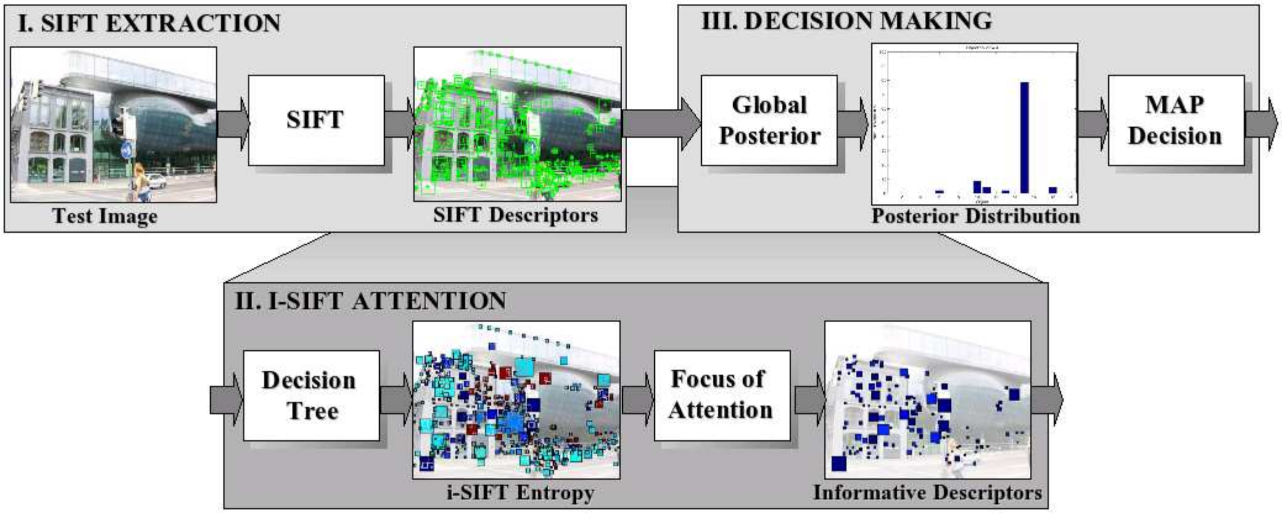


Fig. 3. Concept for recognition from informative local descriptors. (I) First, standard SIFT descriptors are extracted within the test image. (II) Decision making analyzes the descriptor voting for MAP decision. (III) In i-SIFT attentive processing, a decision tree estimates the SIFT specific entropy, and only informative descriptors are attended for decision making (II).

the database [11], [12], despite viewpoint, illumination and scale changes in the object image captures. Therefore SIFT is the choice for implementation in urban environments where illumination and scale changes are usually the cause for degrading performances. [13] proposed the *Informative Features Approach* following previous work on informative patches for recognition [9] by using local density estimations to determine the posterior entropy, making local information content explicit with respect to object discrimination. In contrast to [9], [14] who model mutual information between features and objects, the posterior entropy measure would be tolerant to include features with few occurrences, enabling to represent objects by single images. This approach seems particularly suited for the mobile vision tasks in the proposed application, and if attentive recognition and fast response times are requested under real world conditions.

Informative Descriptors We propose here as innovative step to extend the *Informative Features Approach* [13] to *local descriptors*. From a given descriptor we determine the information content from a posterior distribution with respect to given task specific hypotheses. In contrast to costly *global* optimization, one expects 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 descriptor space. We are primarily interested to get the *information content* of any sample local descriptor $\mathbf{d}_i$ in descriptor space $\mathscr{D}$, $\mathbf{d}_i \in \mathscr{R}^{|\mathscr{D}|}$, with respect to the task of object recognition, where o_i denotes an object hypothesis from a given object set $\mathscr{S}_O$. For this, we need to estimate the entropy $H(O|\mathbf{d}_i)$ of the posterior distribution $P(o_k|\mathbf{d}_i)$, $k = 1\ldots\Omega$, Ω is the number of instantiations of the object class variable O. The Shannon conditional entropy denotes

$$H(O|\mathbf{d}_i) \equiv -\sum_k P(o_k|\mathbf{d}_i) \log P(o_k|\mathbf{d}_i). \qquad (1)$$

One approximates the posteriors at $\mathbf{d}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window of a local neighborhood ε, $||\mathbf{d}_i - \mathbf{d}_j|| \leq \varepsilon$, $j = 1\ldots J$. Fig. 3 depicts *discriminative descriptors* in an entropy-coded representation of local SIFT features $\mathbf{d}_i$. 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 $\mathbf{d}_i$ with $\hat{H}(O|\mathbf{d}_i) \leq H_\Theta$. A specific choice on the threshold H_Θ consequently determines both storage requirements and recognition accuracy (Sec. V). To speed up the matching we use efficient memory indexing of nearest neighbor candidates described by the adaptive *K-d* tree method.

i-SIFT descriptors We apply the *Informative Feature Approach* on Scale Invariant Feature Transform (SIFT [1]) based descriptors that are among the best local descriptors with respect to invariance to illumination changes, matching distinctiveness, image rotation, and blur [12]. The i-SIFT approach tackles three key bottlenecks in SIFT estimation: i-SIFT will (i) improve the recognition accuracy with respect to class membership, iii) provide an entropy sensitive matching method to reject non-informative outliers and more efficiently reject background, (iii) obtain an informative and sparse object representation, reducing the high dimensionality (128 features) of the SIFT keypoint descriptor and thin out the number of training keypoints using posterior entropy thresholding, as follows,

1) *Information theoretic selection of representation candidates*. We exclusively select *informative* SIFT descriptors for object representation. The degree of reduction in the number of training descriptors is determined by threshold H_Θ for accepting sufficiently informative descriptors, practically reducing the representation size by up to one order of magnitude.

2) *Entropy sensitive matching* in nearest neighbor indexing is then necessary as a means to reject outliers in analyzing test images. Any test descriptor $\mathbf{d}_*$ will be rejected from matching if it comes not close enough to any training descriptor $\mathbf{d}_i$, i.e., if $\forall \mathbf{d}_i : |\mathbf{d}_i - \mathbf{d}_*| < \varepsilon$, and ε was determined so as to optimize posterior distributions with respect to overall recognition accuracy.
3) *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 [1] 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, where PCA is applied to the normalized gradient pattern, but that also becomes more errorprone under illumination changes [12].

IV. Attentive Object Detection

i-SIFT based object detection (Sec. III) can achieve a significant speedup from attentive filtering for the rejection of less promising candidate descriptors. This rapid attentive mapping is proposed here 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 for investigation.

Object Detection and Recognition 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 (Fig. 3),

1) **SIFT Extraction** and mapping into PCA based descriptor subspace.
2) **Attentive Mapping** from subspace to an associated estimated entropy value via *decision tree*.
3) **Rejection** of descriptors contributing to ambiguous information (*focus of attention*).
4) **Nearest neighbor analysis** for selected descriptor hypotheses (*global posterior I*).
5) **Posterior estimation** from the histogram of hypothesis specific descriptors (*global posterior II*).
6) **Background rejection** for high entropy posteriors.
7) **MAP classification** for object identifications.

From a given test image, SIFT descriptors are extracted and mapped to an entropy value (see below). An entropy threshold H_Θ for rejecting ambiguous, i.e., high entropy descriptors is most easily identical with the corresponding threshold applied to get a sparse model of reference points (Sec. III). For retained descriptors, we search for the object hypothesis of the nearest neighbor training descriptor. All hypotheses of an image feed into a global histogram which is normalized to give the posterior with respect to object hypotheses. Background rejection is efficiently operated by using a predefined threshold either on the maximum confidence of the MAP hypothesis or the entropy in the posterior.

TABLE I

Confusion map of the C4.5 based decision tree learned on the MPG-20 images (Fig. 4a). Individual entropy intervals – by classes $H_1 \cdots H_5$ – partition $[0, H_{max}]$ into equally large intervals (Sec. V) and are mapped well to output classes.

maps ↦	$\hat{H}_1$	$\hat{H}_2$	$\hat{H}_3$	$\hat{H}_4$	$\hat{H}_5$
H_1	1017	451	197	57	10
H_2	314	1114	196	92	16
H_3	150	185	1171	185	41
H_4	57	125	194	1205	151
H_5	10	15	64	163	1480

Attention using Decision Trees For a rapid estimation of SIFT entropy values, the descriptor attribute values are fed into the decision tree which maps SIFT descriptors $\mathbf{d}_i$ into entropy estimates $\hat{H}$, $\mathbf{d}_i \mapsto \hat{H}(O|\mathbf{d}_i)$. The C4.5 algorithm [15] 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. Table I gives the example of a confusion table that illustrates the quality of mapping *PCA encoded* SIFT descriptors to entropy values. The extraction of informative SIFTs (i.e., i-SIFTS) in the image is performed in two stages (Fig. 3). First, the decision tree based entropy estimator provides a rapid estimate of local information content of a SIFT key under investigation. Only descriptors $\mathbf{d}_i$ with an associated entropy below a predefined threshold $\hat{H}(O|\mathbf{d}_i) < H_\Theta$ are considered for recognition. Only these selected discriminative descriptors are then processed by nearest neighbor analysis with respect to the object model, and interpreted via MAP decision analysis.

Computational Complexity There are several practical issues in using i-SIFT attentive matching that significantly ease the overall 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* (Table II), 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 practically *decreases computational load down to* $<<$ *30% (*$<$ *5% in ZuBuD recognition, Sec. V)*. Finally, the attentive decision tree based mapping is applied to reject SIFT descriptors, thereby *retaining only about* $\leq$ 20% SIFT descriptors for further analysis. These performance differences do hold regardless of using exact (in a k-d tree) or approximate (Best-Bin-First) nearest neighbor search [1].

TABLE II

Object detection results on the MPG-20 mobile imagery (Sec. V), comparing *standard SIFT* keypoint matching [1] and *i-SIFT* attentive matching. i-SIFT provides better MAP accuracy, provides a better detection rate (PT) with less false alarms (PF), and achieves significantly better avg. posterior entropy H and avg. MAP confidences.

recognition method	MAP accuracy MPG-20 [%]	PT [%]	PF [%]	obj avg. H	bgd avg. 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

TABLE III

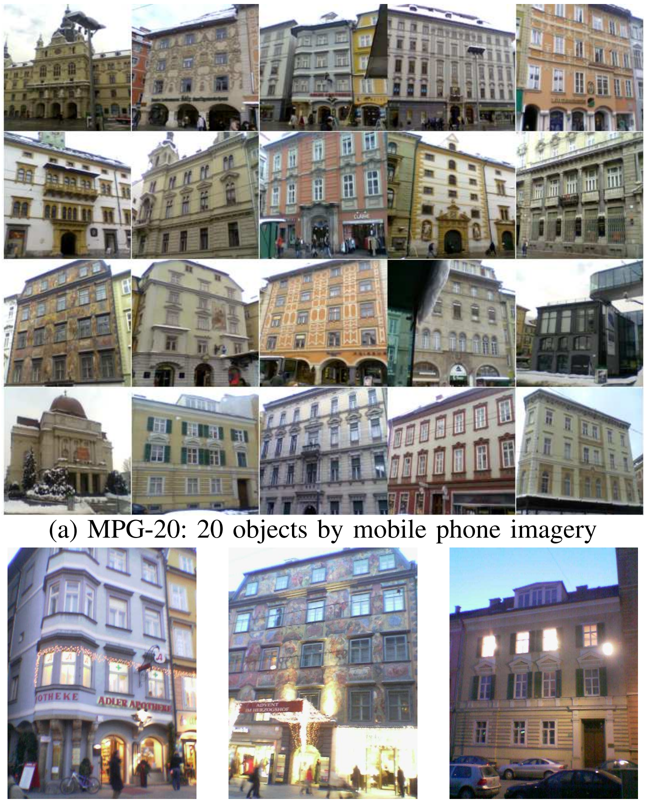
Runtime performance results for the MPG-20 experiments. The i-SIFT method (modules M1: PCA projection, M2: decision tree, M3: nearest-neighbor search) required on avg. $\approx 2.8sec$ per image for object identification, being ≈ 8 times faster than SIFT recognition, requiring only $\approx 12\%$ storage.

recognition method	descriptors	recognition stages			**total**	no. keys
SIFT	1.8 sec	7.48 sec (ratio method)			**9.28 sec**	28873
i-SIFT	1.8 sec	0.08 sec (M1)	0.01 sec (M2)	0.91 sec (M3)	**2.80 sec**	3501

V. Experiments

Targeting emerging technology applications using computer vision on mobile devices, we perform the performance tests using the i-SIFT approach on mobile phone imagery captured about tourist sights in the urban environment of the city of Graz, Austria, i.e., from the MPG-20 database (Fig. 4a), and illustrate performance improvements gained from the i-SIFT approach in comparison to standard SIFT matching. We present results proving a *reliable mobile vision service* for urban object detection.

MPG-20 Database The MPG-20 database[1] includes images about 20 objects, i.e., front sides of buildings from the city of Graz, Austria, captured in a user test trial by students. Most of these images contain a tourist sight, some containing non-planar structure (o_3, o_5, o_{16}, Fig. 4a), together with 'background' information from surrounding buildings, pedestrians, etc. The images of 640×480 pixels were captured from an off-the-shelf camera-equipped phone (Nokia 6230), containing 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 and date. For each object, we then selected 2 images taken by a viewpoint change of $\approx \pm 30^\circ$ and of 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 significant scale change - were taken for test purposes, giving 40 test images in total. Additional test images were obtained (i) from other 'non-sight' buildings and natural landscapes which are not part of MPG-20, i.e., 'background', and (ii) about MPG-20 objects under extreme illumination conditions (e.g., in the evening twilight, Fig. 4b).



(a) MPG-20: 20 objects by mobile phone imagery

(b) eMPG-20: 20 objects under extreme illumination

Fig. 4. 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. V).

SIFT based Key Matching The grayvalued training images (colour was assumed too sensitive to illumination changes) were bit-masked by hand, such that SIFT descriptors on background information (surrounding 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. Object recognition is then performed using MAP decision making (Sec. IV). 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

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

confidence was used as decision criterion to discriminate between object ($> 25\%$) and background ($\leq 25\%$) images (for both SIFT and i-SIFT, see Fig. 6). For the training of the **i-SIFT** selection, the descriptors were first projected to an eigenspace of dimension 40, thereby decreasing the original descriptor input dimensionality (128 features) by a factor of three. A decision tree [15] of depth 52 was learned for the attentive matching, defining the threshold for attentive matching by $H \leq H_\Theta = 1.0$. 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 97.5% (SIFT: 95%), the average entropy in the posterior distribution was $H_{avg} \approx 0.5$.

MPG-20 Performance Results Table II depicts results of the MPG-20 experiments, and comparing SIFT vs. i-SIFT keypoint matching. i-SIFT provides better MAP accuracy, better detection rate (PT) with less false alarms than using SIFT, being able to provide robust discrimination between object and background images, by using a MAP confidence threshold to accept/reject object hypotheses. The runtime for single image recognition (PC Pentium IV, 2.4 GHz, C++ non-optimized code) was 2.80*sec* using i-SIFT (in contrast to 9.82*sec* with standard SIFT), demonstrating that i-SIFT should be preferred for mobile object awareness (Table III). Note that SIFT based recognition with i-SIFT-like model complexity (retaining only 12% of SIFT training keys by random selection) decreases to 32.5% – *i-SIFT is truly informative*! An important issue for mobile services represents the guarantee for reliable quality of service. From the challenging experiments with 80 object images under extreme illumination conditions (Fig. 4b), we finally derived a threshold on the minimum number of keypoint votes ($>> 4$) required for detection decisions to be communicated to a user (otherwise, the system would inform uncertain conditions). Based on this threshold the system provided a trustworthy mobile service, achieving 100% accuracy – even under the extreme conditions reported in Fig. 4b – for accepted (50%) object images, rejecting (50%) for annotation otherwise.

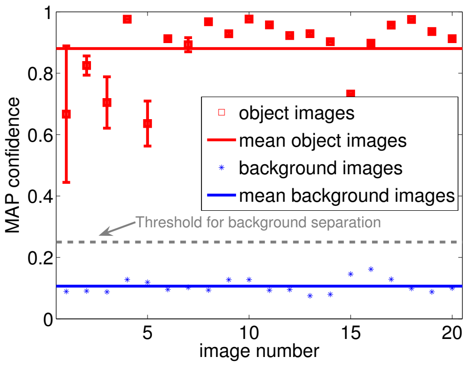
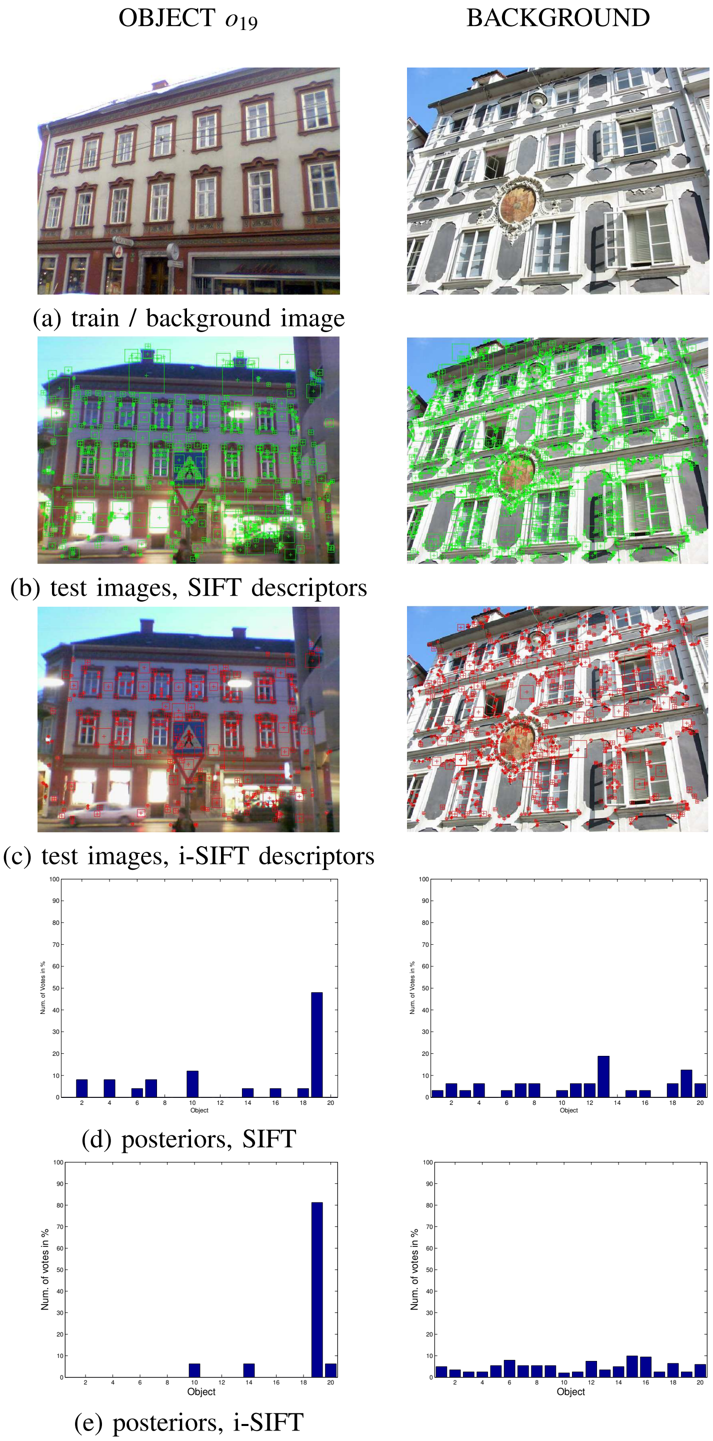


Fig. 5. Performance results on MPG-20 imagery using i-SIFT supported MAP confidence based discrimination between object and background.



(a) train / background image

(b) test images, SIFT descriptors

(c) test images, i-SIFT descriptors

(d) posteriors, SIFT

(e) posteriors, i-SIFT

Fig. 6. Sample object detection results for object object o_{19} (left) and background (right). (a) Depicting train and bgd. images. (b) SIFT descriptor locations on test images. (c) Selected i-SIFT descriptor locations. (d) Posterior distribution on object hypotheses from SIFT and (e) i-SIFT descriptors, respectively, demonstrating more discriminative results for i-SIFT based interpretation.

ZuBuD Performance Results In similar manner, we applied i-SIFT key matching to the ZuBuD database[2] (201 buildings, 5 views each, 115 query images [4]). While [4] achieved only moderate performance ($\approx 86\%$ accuracy), our i-SIFT system achieved $\approx 91\%$ correct identifications. Note that the avg. runtime per image for i-SIFT based recognition was $4.8 sec$ in contrast to $115 sec$ by standard SIFT (due to the large search list), making i-SIFT ≈ 24 times faster than SIFT !

VI. Summary and Conclusions

A methodology for reliable urban object detection was presented using off-the-shelf camera-equipped mobile devices. The *Informative Descriptor Approach* was applied to SIFT keys, resulting in significant performance improvements in object detection, with respect to detection rates, efficient use of memory resources and speedup in the recognition process. This paper also introduced *attentive matching* for descriptors, applying an information theoretic criterion for the selection of discriminative SIFT descriptors for recognition matching and representation. This innovative local descriptor is most appropriate for sensitive operation under limited resources, such as, in applications using mobile vision services. We evaluated the detection system on the public available MPG-20 database, including images from 20 building objects, 'non-object' images, and extreme illumination conditions about Graz urban environment.

The proposed urban object detection system could have a strong impact on many areas of m-commerce, such as, tourist information systems, navigation aids for the visually impaired, mobile learning, mobile inspection, etc. Future work goes in the direction of exploiting geometric relations between informative descriptors to provide robust grouping and segmentation of categorical object information.

[2] http://www.vision.ee.ethz.ch/showroom/zubud/index.en.html

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Y. Li and L. Shapiro, "Consistent line clusters for building recognition in CBIR," in *Proc. International Conference on Pattern Recognition, ICPR 2002*, vol. 3, 2002, pp. 952–957.

[3] A. Dick, P. Torr, and R. Cipolla, "Modelling and interpretation of architecture from several images," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 111–134, 2004.

[4] H. Shao, T. Svoboda, and L. van Gool, "HPAT indexing for fast object/scene recognition based on local appearance," in *Proc. International Conference on Image and Video Retrieval, CIVR 2003*. Chicago,IL, 2003, pp. 71–80.

[5] T. Yeh, K. Tollmar, and T. Darrell, "Searching the web with mobile images for location recognition," in *Proc. IEEE Computer Vision and Pattern Recognition, CVPR*, Washington, DC, 2004, pp. 76–81.

[6] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins, *Global Positioning System Theory and Practice*. Vienna, Austria: Springer-Verlag, 2001.

[7] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. European Conference on Computer Vision, ECCV 2002*, 2002, pp. 128–142.

[8] S. Obdrzalek and J. Matas, "Object recognition using local affine frames on distinguished regions," in *Proc. British Machine Vision Conference*, 2002, pp. 113–122.

[9] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classication," in *Proc. International Conference on Computer Vision, ICCV 2003*. Nice, France, 2003, pp. 281–288.

[10] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2003, pp. 264–271.

[11] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," in *Proc. Computer Vision and Pattern Recognition, CVPR 2003*, Madison, WI, 2003.

[12] ——, "A performance evaluation of local descriptors," http://www.robots.ox.ac.uk/ vgg/research/affine/, 2004.

[13] G. Fritz, L. Paletta, and H. Bischof, "Object recognition using local information content," in *Proc. International Conference on Pattern Recognition, ICPR*, vol. II. Cambridge, UK, 2004, pp. 15–18.

[14] G. Dorko and C. Schmid, "Selection of scale-invartiant parts for object class recognition," in *Proc. International Conference on Computer Vision, ICCV 2003*, 2003, pp. 634–640.

[15] J. Quinlan, *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.