Learning Informative SIFT Descriptors for Attentive Object Recognition

Christin Seifert [1], *Gerald Fritz* [1], *Lucas Paletta* [1] *and Horst Bischof* [2]

[1] *JOANNEUM RESEARCH - Institute of Digital Image Processing*
Wastiangasse 6, A-8010 Graz, Austria
{*christin.seifert,gerald.fritz,lucas.paletta*}*@joanneum.at*

[2] *Graz University of Technology - Institute for Computer Graphics and Vision*
Inffeldgasse 16/II, A-8010 Graz, Austria
bischof@icg.tu-graz.ac.at

Abstract:
With the emerging sensor technologies in mobile devices, such as camera phones, visual interpretation methodologies are challenged to provide solutions within the everydays outdoor urban environment. For this purpose, we propose to apply the 'Informative Descriptor Approach' on the SIFT descriptor [4], in order to define the informative SIFT (i-SIFT) descriptor. By attentive matching of i-SIFT keypoints, we provide an innovative method on object detection that significantly improves SIFT based keypoint matching. i-SIFT tackles the SIFT bottlenecks, e.g., extensive nearest neighbor indexing, by (i) significantly reducing the descriptor dimensionality, (ii) decreasing the size of object representation by one order of magnitude, and (iii) performing matching exclusively on attended descriptors, as required by resource sensitive devices. The key advantages of informative SIFT (i-SIFT) are demonstrated in a typical outdoor mobile vision experiment on the TSG-20 reference database, detecting buildings with high accuracy.

1 Introduction

Research on visual object recognition and detection has recently focused on the development of local interest operators [1, 4, 3] and the integration of local information into robust object recognition. 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. This methodology 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.

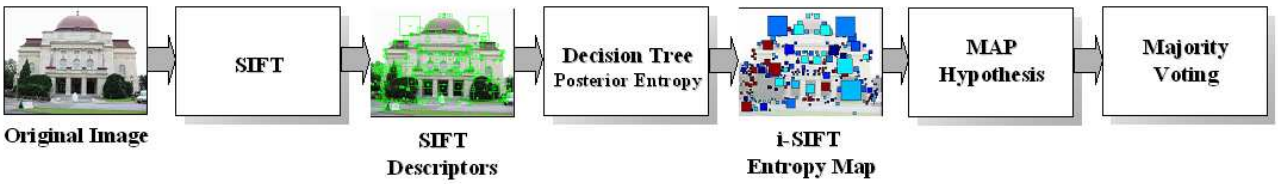


Figure 1: Automated building recognition. Standard SIFT descriptors are first extracted within the test image. The proposed informative SIFT (i-SIFT) approach determines the entropy in the descriptor and performs decision making (MAP hypothesizing) only on attended descriptors. Majority voting is then used to integrate local votes into a global classification.

The key contribution of the presented work is to extend the Informative Descriptor Approach [2] to complex features to perform recognition in outdoor urban environments more efficient (Fig. 1). We intend to extract informative features of the state-of-the-art SIFT interest point detector [4], which is known to be less sensitive to scale and illumination changes compared to other local descriptors [6]. The few drawbacks in SIFT based recognition are identified as (i) the computational complexity in nearest neighbor indexing, and (ii) lack in representation of uncertainty. We tackle these issues by (i) matching only informative keypoints (*attentive matching*) and (ii) estimate posterior distributions on object hypotheses. We use local density estimations to determine conditional entropy measures, and 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.

The experiments were performed on mobile 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 the significant speedup of the algorithm by one order of magnitude, and requiring a fraction ($\approx 20\%$) of features of lower dimensionality (30%) for representation.

2 Detection from Informative Features

The *Informative Feature Approach* requires to extract relevant features in a pre-processing stage by optimizing feature selection with respect to the information content in the context of a specific task, e.g., object recognition. We motivate informative descriptors from information theory and describe the application to local SIFT descriptors.

2.1 Local Information Content

We determine 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 [2].

The object recognition task is applied to sample local descriptors $\mathbf{f}_i$ in feature space $\mathcal{F}$, $\mathbf{f}_i \in \mathcal{R}^{|\mathcal{F}|}$, where o_i denotes an object hypothesis from a given object set Ω. We need to estimate the entropy $H(O|\mathbf{f}_i)$ of the posteriors $P(o_k|\mathbf{f}_i)$, $k = 1 \ldots \Omega$, Ω is the number of instantiations of the object class variable O. Shannon conditional entropy denotes

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

We approximate the posteriors at $\mathbf{f}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window of a local neighborhood ϵ,

$$||\mathbf{f}_i - \mathbf{f}_j|| \leq \epsilon, \tag{2}$$

$j = 1 \ldots J$. 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 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 $\mathbf{f}_i$.

We receive sparse instead of extensive object representations, in case we store only *selected* imagette information that is *relevant for classification* purposes, i.e., *discriminative* $\mathbf{f}_i$ with $\hat{H}(O|\mathbf{f}_i) \leq \Theta$. A specific choice on the threshold Θ consequently determines both storage requirements and recognition accuracy. For efficient memory indexing of nearest neighbor candidates we use the adaptive K-d tree method.

2.2 Informative SIFT Descriptors

[5, 6] give a thorough performance comparison on descriptors of local interest regions, concluding that the SIFT based descriptors perform mostly best, with respect to matching distinctiveness, invariance to blur, image rotation, and illumination changes. 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 training keypoints to obtain an informative, sparse object representation, and (iii) providing an entropy sensitive matching method to reject non-informative outliers, as follows,

1. *Reduction of feature dimensionality* (128 features) of the SIFT descriptor is crucial to keep nearest neighbor indexing computationally feasible. To discard statistically irrelevant feature dimensions, we applied Principal Component Analysis (PCA) on the SIFT descriptors, in contrast to the PCA-SIFT method [3], where PCA is applied to the gradient pattern (more errorprone under illumination changes [6]).

2. *Information theoretic selection of representation candidates.* 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. 3) this approximately reduces the representation size by one order of magnitude.

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.

2.3 Object Recognition and Detection

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.

3. **Rejection** of imagettes 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 $o_i \in O$. In case this descriptor would convey ambiguous information, it is removed from further consideration, getting as well sparse models of reference points. Object recognition on a collection of (matched and therefore labelled) SIFT descriptors is then performed by majority voting on the complete set of class labels attained from individual descriptor interpretations.

Attentive Matching For a rapid estimation of local entropy quantities, the descriptor encoding is fed into a 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 [7] 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.

Rapid extraction of i-SIFTs 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.

i-SIFT provides improvements in computational load 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.

3 Experimental results

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. Targeting emerging technology applications using computer vision on mobile devices, we perform the performance tests on mobile imagery captured about tourist sights in the urban environment of the city of Graz, Austria, i.e., from the TSG-20 database (see below, Fig. 2).

TSG-20 Database The TSG-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 contain 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. The images were first subsampled to size 240×320. For each object, we then selected 2 images taken by a viewpoint change of $\approx \pm 30^\circ$ of a similar distance to the

[1)] The TSG-20 (Tourist Sights Graz) database can be downloaded at the URL http://dib.joanneum.at/cape/TSG-20.

Method	Acc. [%]	NN [%]	$\bar{H}$	Pts.	dim
i-SIFT,Θ=1.8	100.0	39.8	2.14	178	40
i-SIFT,Θ=1.0	97.5	23,1	1.45	178	40
SIFT	97.5	100,0	1.35	711	128

Table 1: Performance comparison between *i-SIFT* attentive matching and *standard SIFT* keypoint matching [4] on TSG-20 mobile imagery (Sec. 3, acc.=accuracy, NN=nearest neighbor searches, $\bar{H}$=average entropy measure, Pts.= training descriptors, and dim=dimensionality of descriptors).

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.

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 0.45, and 178 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 7125, i.e., $\approx 31\%$ 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 $H_{avg} \approx 1.4$ (as in SIFT approach).

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

4 Conclusions

The presented work aimed at outdoor building recognition, applying the *Informative Descriptor Approach* to complex local descriptors, significantly improving the efficiency in object detection, both with respect to memory resources and to speedup the recognition process, using an improvement of the SIFT approach [4] by i-SIFT entropy thresholded descriptor matching.

The i-SIFT 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 TSG-20 database, including images from 20 building objects and 'non-object background' pictures from the city of Graz, comparing well with SIFT high recognition accuracy while providing posterior distributions for reasoning, robust background detection, and significant speedup in processing times. Future work goes in the direction of exploiting ge-

ometric relations between i-SIFT features to provide robust grouping and segmentation of object specific information.

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] 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.

[3] 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.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

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

[6] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. http://www.robots.ox.ac.uk/ vgg/research/affine/, 2004.

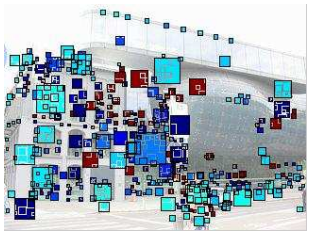
[7] J.R. Quinlan. *C4.5 Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

Object o_{15}

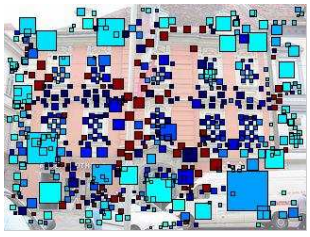
Object o_8

BGD o_{22}

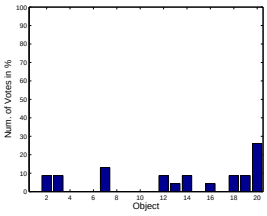
BGD o_{24}



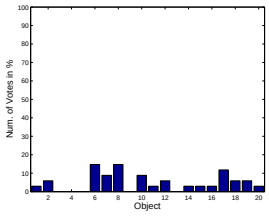
entropy in SIFTs – o_{15}



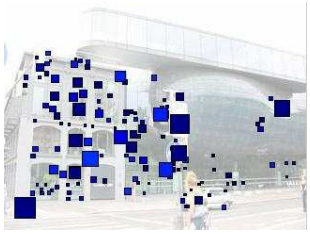
entropy in SIFTs – o_8



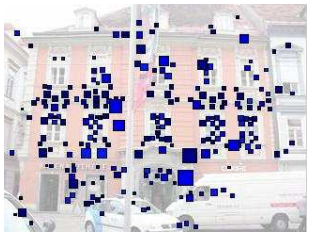
posterior i-SIFT o_{22}



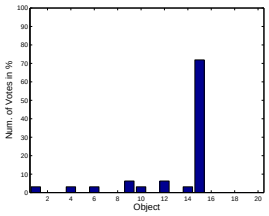
posterior i-SIFT o_{24}



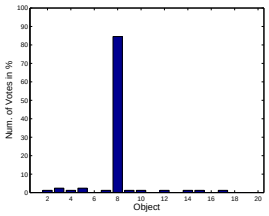
attended i-SIFTs for o_{15}



attended i-SIFTs for o_8



posterior i-SIFT o_{15}



posterior i-SIFT o_8

Figure 2: Left 2 columns: Sample building recognition for objects o_{15}, o_8 (top-down): train images, entropy coded (blue:low, red: high) SIFT descriptors (without selection), entropy coded and *attended*, object specific i-SIFT descriptors, and corresponding posterior distribution for i-SIFT based descriptor recognition. Right 2 columns: detection of background by high entropy (low confidences) in the posteriors.