Geo-Indexed Object Recognition for Mobile Vision Tasks

Katrin Amlacher
katrin.amlacher@joanneum.at

Lucas Paletta
lucas.paletta@joanneum.at

JOANNEUM RESEARCH Forschungsgesellschaft mbH
Institute of Digital Image Processing
Wastiangasse 6, 8010 Graz, Austria

ABSTRACT

The presented work settles attention in the architecture of ambient intelligence, in particular, for the application of mobile vision tasks in multimodal interfaces. A major issue for the performance of these services is uncertainty in the visual information which roots in the requirement to index into a huge amount of reference images. The presented functional component – the *Attentive Machine Interface* (AMI) – enables contextual processing of multi-sensor information in a probabilistic framework, for example to exploit contextual information from geo-services with the purpose to cut down the visual search space into a subset of relevant object hypotheses. We demonstrate results about geo-indexed object recognition from experimental tracks and image captures in an urban scenario, extracting object hypotheses in the local context from both (i) mobile image based appearance and (ii) GPS based positioning, and verify performance in recognition accuracy ($> 14\%$) using Bayesian decision fusion, verifying the advantage of multi-sensor attentive processing in multimodal interfaces.

Categories and Subject Descriptors

I.1.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*; G.3 [**Probability and Statistics**]: Statistical computing

General Terms

Algorithms, Performance

1. INTRODUCTION

Attention as a methodology of selecting detail of relevance is ubiquitious in biological systems and has increasingly received consideration for the design of artificial cognitive systems. In this paper we approach attention from the viewpoint of a nomadic urban user that is equipped with a camera phone and that is interested in receiving appropriate information about objects of interest within a local environment. We describe the embedding of the problem in a general system implementation of an *Attentive Machine Interface* (AMI) that enables contextual processing of multi-sensor information in a probabilistic framework. The system is prepared to support *bottom-up* information processing in terms of selecting and processing information within specific modalities and according to a pre-defined – be it learned or heuristically determined – methodology. A particularly novel functionality presented in this work is to enable *top-down* information processing by cross-modal priming of early processing in the manner of a multi-sensor framework for attentive – and finally superior – performance.

In ubiquitous computing, several frameworks have been proposed in the frame of attentive interfaces and context awareness. [8] proposed Attentive User Interfaces that capture the attention of the user, e.g. from eye gaze estimation, and consequently adapt interaction systems for better communication with the user. [1] proposed that context is a description of a real world situation on an abstract level that is derived from available cues. Finally, [6] proposed a context processing system with blackboard functionality where components can subscribe to receive messages matching specific patterns, and various cues are integrated into a multimodal description of a situation. While the concept of AMI is directly inspired by [6], it presents processing in a probabilistic framework and enables top-down, i.e., attentive cross-modal information processing.

Mobile object recognition and visual positioning have recently been proposed in terms of mobile vision services for the support of urban nomadic users. A major issue for the performance of these services is uncertainty in visual information; covering large urban areas with naive approaches would require to refer to a huge amount of reference images and consequently to highly ambiguous features. Previous work on mobile vision services primarily advanced the state-of-the-art in computer vision methodology for the application in urban scenarios. [7] provided a first innovative attempt on building identification proposing local affine features for object matching. [9] introduced image retrieval methodology for the indexing of visually relevant information from the web for mobile location recognition. Subsequent attempts (e.g., [5, 2]) advanced the methodology further towards higly robust building recognition. However, so far it has not been considered to investigate the contribution of geo-information to the performance of the vision service.

We propose to exploit contextual information from geo-services with the purpose to cut down the visual search space into a subset of all available object hypotheses in the large urban area. Geo-information in association with visual features enables to restrict the search within a local context. The results from experimental tracks and image captures in an urban scenario prove a significant increase in recognition accuracy (Sec. 4) and use of computational resources when using in contrast to omitting geo-contextual information.

MobileHCI 2008, September 2-5, 2008, Amsterdam, the Netherlands.

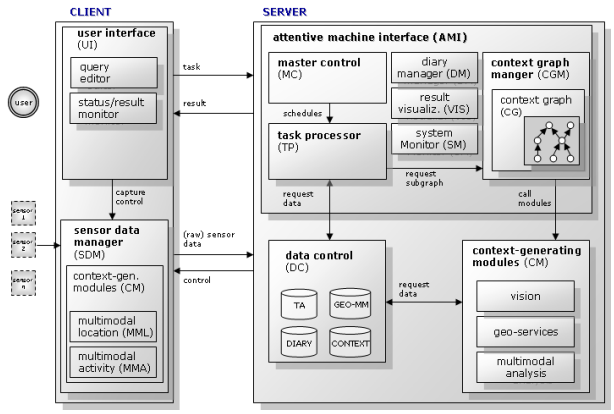

Figure 1: Concept of a client-server system architecture with Attentive Machine Interface (AMI).

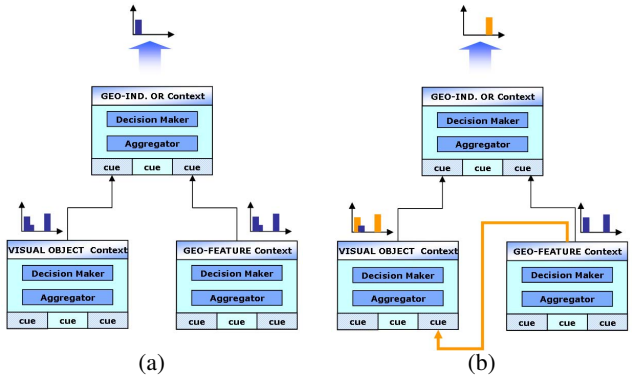

Figure 2: Context service "Geo-Indexed Building Recognition" and context nodes for (a) bottom-up information processing of visual object recognition and geo-features, and for (b) top-down information processing using geo-features to prime recognition.

2. ATTENTIVE MACHINE INTERFACE

The context framework used in the AMI defines a cue as an abstraction of logical and physical sensors which may represent a context itself, generating a recursive definition of context. Sensor data, cues and context descriptions are defined in a framework of uncertainty. Attention is the act of selecting and enabling detail – in response to situation specific data - within a choice of given information sources, with the purpose to operate exclusively on it. Attention enabled by the AMI is therefore focusing operations on a specific detail of a situation that is described by the context.

Fig.. 1 outlines the embedding of the AMI within a client-server system architecture for mobile vision services with support from multi-sensor information. A user interface generates task information (mobile vision service) that is fed into the Master Control (MC) and then the Task Processor (TP) who, firstly, requests a hierarchical description of services, i.e. context-generating modules (context subgraph) and, secondly, executes the services in the order of the subgraph description. Context-generating modules are services that receive an input cue (an image, a GPS signal, etc.) from the Data Control (DC) module and generate a specific context abstraction from an integration of the input cues.

Context Processing For the generation of high-level context information only parts of the Context Graph need to be processed, in fact those that contribute to the corresponding (top-level) context node. Depending on available input data and services, a subgraph from the Context Graph is derived which consequently ensures a smooth processing by the Task Processor. The subgraph gets processed starting with those leaf context nodes which take data only from the Data Control. The calculated results are given to the next Context Nodes following the outgoing edges until the top-level context node is reached. Resulting high-level context information is given to the user via a visualization compent and is stored in the Data Control or Diary Manager.

Bottom-Up and Top-Down Processing The AMI supports two different modes of information processing, i.e., bottom-up and top-down processing. The choice of modes can be decided by the Task Processor according to demands on computational resources, quality of service (e.g., recognition accuracy) and availability of data. Figure 2 provides a schematic sketch of the service of *geo-indexed object recognition* (Sec. 3) in both processing modes. In bottom-up processing mode (a), services for the computation of (i) visual objects (object recognition) and (ii) geo-features (positioning) are determining hypotheses with respect to the occurrence of objects (i) in the image and (ii) within a local environment. In top-down processing mode (b), there is a cross-modal dependency in (i) object recognition on the input of object hypotheses provided by (ii) the geo-service. While individually processed distributions on object hypotheses can simply be integrated in (a) using Bayesian decision fusion, (b) actually models an impact of geo-information on visual feature extraction as outlined in more detail in Sec. 3.

3. GEO-INDEXED RECOGNITION

Urban image based recognition provides the technology for both situated object awareness and positioning. We briefly describe how we make use of the methodology presented in [2]. The user captures an image about an object of interest in its field of view, and a software client initiates wireless data submission to the server. Assuming that a GPS receiver is available, the mobile device reads the actual position estimate and sends this together with the image to the server. In the second stage, 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.

Attentive Object Recognition Research on visual object detection has recently focused on the development of local interest operators (e.g., [3]) and the integration of local information into object recognition. The *Informative SIFT Features Approach* (i-SIFT [2]) provides robust matching despite viewpoint, illumination and scale changes in the object image captures which is mandatory for mobile vision services. It uses local density estimations in feature space to determine the posterior 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, where o_i denotes an object hypothesis from a given object set $\mathcal{S}_O$. .

Fig. 3 depicts *discriminative descriptors* in an entropy-coded representation of local SIFT features $\mathbf{d}_i$. From discriminative local descriptors one proceeds 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. For the rejection of images whenever they do not contain any objects of interest one estimates the entropy in the posterior distribution and rejects images with posterior entropies above a certain threshold.

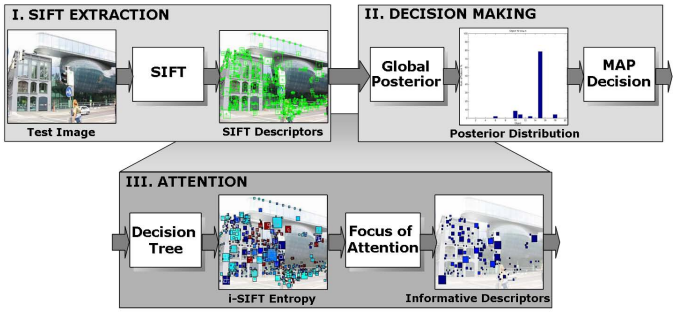

Figure 3: Concept for recognition from informative local descriptors. (I) SIFT descriptors are extracted within the test image. (II) Decision making analyzes the descriptor voting for MAP decision. (III) In i-SIFT, a decision tree exclusively selects informative descriptors for decision making (II).

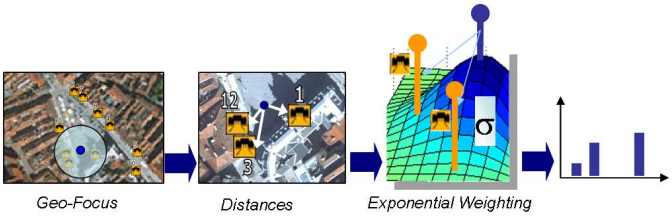

Figure 4: Extraction of object hypotheses from geo-services. (Left to right) Within a local spatial neighborhood (geo-focus), distances to the points of interest are exponentially weighted and normalised to provide a distribution on object hypotheses.

Geo-Contextual Computing of Object Recognition Geo-services provide access to information about a local context that is stored in a digital city map. Map information in terms of map features is indexed via a current estimate on the user position that can be derived from satellite based signals (GPS), dead-reckoning devices and so on. The map features can provide geo-contextual information in terms of, e.g., location of points of interest. In previous work [4], the general relevance of geo-services for the application of mobile object recognition was already emphasised, however, the contribution of the geo-services to the performance of geo-indexed object recognition was not quantitatively assessed, and top-down processing was not considered.

Fig. 4 depicts a novel methodology to introduce geo-service based object hypotheses. (i) A geo-focus is first defined with respect to a radius of expected position accuracy with respect to the city map. (ii) Distances between user position and points of interest (e.g., tourist sight buildings) that are within the geo-focus are estimated. (iii) The distances are then weighted according to a normal density function by $p(\mathbf{x}) = 1/((2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}) \exp\{-1/2(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\}$. By investigating different values for σ, assuming $(\mathbf{\Sigma}_{ij}) = \delta_{ij}\sigma_j^2$, one can tune the impact of distances on the weighting of object hypotheses. (iv) Finally, weighted distances are normalised and determine confidence values of individual object hypotheses.

Bottom-Up Geo-Indexed Object Recognition Distributions over object hypotheses from vision and geo-services are then integrated via Bayesian decision fusion. Although an analytic investigation of both visual and position signal based information should prove statistical dependency between the corresponding random variables, one assumes that it is here sufficient to pursue a naive Bayes approach for the integration of the hypotheses (in order to get a rapid estimate about the contribution of geo-services to mobile vision services) by $P(o_k|\mathbf{y}_{i,v}, \mathbf{x}_{i,g}) = p(o_k|\mathbf{y}_{i,v})p(o_k|\mathbf{x}_{i,g})$, where indices v and g mark information from image ($\mathbf{y}$) and positioning ($\mathbf{x}$), respectively.

Top-Down Geo-Indexed Object Recognition Here, we firstly process the geo-service in order to receive a distribution over object hypotheses that is input to attentive object recognition. The recognition method is then primed to reject all those local (i-SIFT; see above) descriptors from consideration that are labelled with hypotheses of negligible confidence in the output of the geo-service. Hence the feature space underlying the nearest-neighbor voting procedure is containing only pre-selected prototypes which are then preferred but outside a pre-determined distance threshold in feature space. The resulting distribution over object hypothesis can again be fused with the distribution from geo-services in order to recieve a distance based weighting on object hypotheses.

4. EXPERIMENTS

The overall goal of the experiments was to determine and to quantify the contribution of geo-services to object recognition in urban environments and to compare bottom-up and top-down processing modes. The performance in the detection and recognition of objects of interest on the query images with respect to a given reference image database and a given methodology (TSG-20 [2]) was compared to the identical processing but using geo-information and information fusion for the integration of object hypotheses.

User Scenario and Constraints In the application scenario, we imagine a tourist being equipped with a mobile device with built-in GPS. He can send image based queries to a server using UMTS or WLAN based connectivity. The server performs geo-indexed object recognition and is expected to respond with tourist relevant annotation if a point of interest was identified. Reference imagery [2] with 640×480 resolution about building objects of the TSG-20 database[1] were captured from a camera-equipped mobile phone (Nokia 6230), containing changes in 3D viewpoint, partial occlusions, scale changes by varying distances for exposure, and various illumination changes. For each object we 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. A test data set was acquired with an ultra-mobile PC with 1.3 MPixels image captures, with seven images per TSG-20 object from different view points; images were captured on different days under different weather conditions.

Attentive Object Recognition In the first evaluation stage, each individual image query was evaluated for vision based object detection and recognition, then regarding extraction of geo-service based object hypotheses, and finally with respect to Bayesian decision fusion on the individual probability distributions (Sec. 3). Detection is an important pre-processing step to recognition, e.g., to avoid geo-services to support confidences for objects that are not in the query image. Experiments on imagery including background data resulted in a PT rate of 89.2% and a FP rate of 20.1%, probably due to the bad sensor quality. However, once a query image is attributed to the object category, the geo-indexed object recognition will boost the performance in finding more correct hypotheses than using vision alone.

Fig. 5 depicts sample query images associated with corresponding distributions on object hypotheses from vision, geo-services, and using information fusion. The results demonstrate significant

[1] http://dib.joanneum.at/cape/TSG-20/

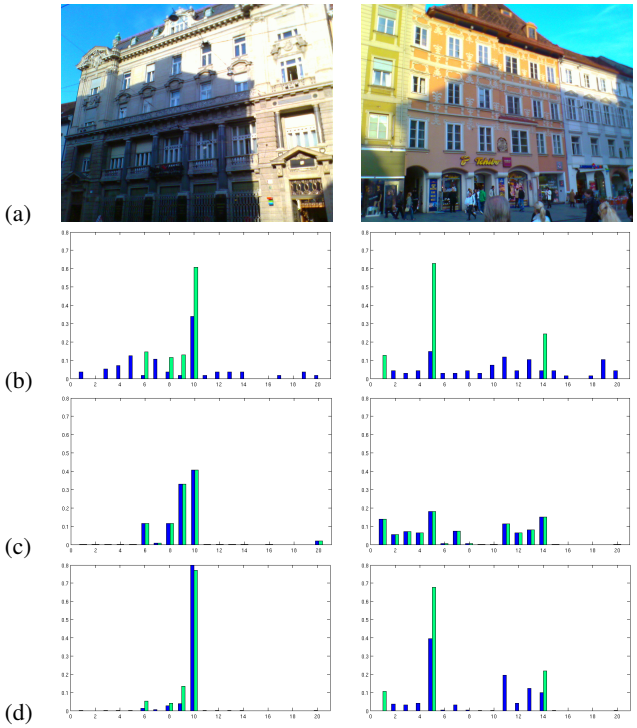

Figure 5: Comparison between bottom-up (blue/dark bars) and top-down approach (green/light bars) from (a) sample input images. Integration of object hypotheses from (b) vision and (c) geo-services into a (d) fused distribution demonstrates clear increases in the confidences of the correct object hypothesis and therefore a significant improvement in the performance of the mobile vision service (Fig. 6).

increases in the confidences of correct object hypotheses. The evaluation of the *complete* database of image queries about TSG-20 objects (Fig. 6) proves a decisive advantage for taking geo-service based information into account in contrast to purely vision based object recognition, in particular, using the top-down approach. While vision based recognition is on a low level ($\approx 84\%$), an exponentially weighted spatial enlargement of the scope on object hypotheses with geo-services increased the recognition accuracy up to $\approx 99\%$. With increasing σ an increasing number of object hypotheses are taken into account for information fusion and the performance finally drops to vision based recognition performance (uniform distribution in the geo-service based object hypotheses).

5. CONCLUSION

In this work we propose the AMI that enables bottom-up and top-down cross-modal information processing. We take advantage of geo-contextual information for the improvement of mobile vision services in urban scenarios, such as visual object recognition of tourist sights. We argued that geo-information provides a focus on the local object context that enables a meaningful selection of expected object hypotheses and therefore improves overall performance of urban object recognition. We proposed to pursue a methodology on Bayesian decision fusion that integrates distributions on object hypotheses from both cues, i.e., visual information and position estimate. We performed experiments on a representative image data set and proved a significant improvement in performance when using geo-services.

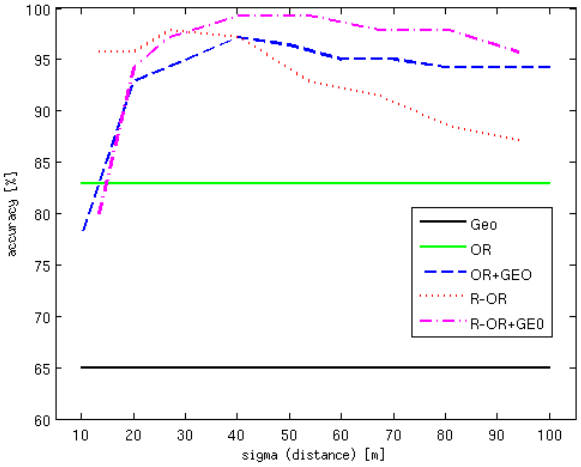

Figure 6: Performance comparison between geo-service based hypotheses (Geo), purely vision based recognition (OR), bottom-up processing with information fusion (OR+GEO), top-down processing of attentive recognition without (R+OR) and with post-processing using Bayesian decision fusion (R+OR+GEO; best results).

In future work we will further exploit the concept of the AMI by integrating different context information, such as visual context or semantic segmentation, in a probabilistic framework.

6. ACKNOWLEDGMENTS

This work is supported in part by the European Commission funded project MOBVIS under grant number FP6-511051 and by the FWF Austrian National Research Network on Cognitive Vision under sub-project S9104-N04.

7. REFERENCES

[1] A. K. Dey and G. D. Abowd. Towards a Better Understanding of Context and Context-Awareness. In *Proceedings of the CHI 2000 Workshop on "The What, Who, Where, When, Why and How of Context-Awareness"*, 2000.

[2] G. Fritz, C. Seifert, and L. Paletta. A Mobile Vision System for Urban Object Detection with Informative Local Descriptors. In *Proc. IEEE 4th International Conference on Computer Vision Systems, ICVS*, New York, NY, January 2006.

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

[4] P. Luley, L. Paletta, A. Almer, M. Schardt, and J. Ringert. Geo-services and computer vision for object awareness in mobile system applications. In *Proc. 3rd Symposium on LBS and Cartography*, pages 61–64. Springer, 2005.

[5] S. Obdrzalek and J. Matas. Sub-linear indexing for large scale object recognition. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 1–10, 2005.

[6] A. Schmidt and K. V. Laerhoven. How to build smart appliances. *IEEE Personal Communications*, pages 66 – 71, 2001.

[7] 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*, pages 71–80. Chicago,IL, 2003.

[8] R. Vertegaal. Attentive User Interfaces. *Communications of the ACM*, 46(3):30–33, 2003.

[9] 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 2004*, pages 76–81, Washington, DC, 2004.