ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-601

An Attentive Machine Interface Using Geo-Contextual Awareness for Mobile Vision Tasks

Katrin Amlacher and **Lucas Paletta** [1]

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. We propose a system implementation – the *Attentive Machine Interface* (AMI) – that 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 present a proof-of-concept with results from bottom-up information processing from experimental tracks and image capture 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 ($> 10\%$) using Bayesian decision fusion. Finally, we demonstrate that top-down information processing – geo-information priming the recognition method in feature space – can yield even better results ($> 13\%$) and more economic computing, verifying the advantage of multi-sensor attentive processing in multimodal interfaces.

1 INTRODUCTION

Attention as a methodology of selecting detail of relevance is ubiquitous in biological systems and has increasingly received consideration for the design of artificial cognitive systems. Mobile multimodal interfaces as devices that receive a multitude and diversity of data with the purpose to assisting the user with relevant detail and level of abstraction are an obvious choice of investigation about how concepts for the appropriate selection of information might contribute to solve a task.

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 in general *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.

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. 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. We extract object hypotheses in the local context from (i) mobile image based appearance and (ii) GPS based positioning and investigate the performance of Bayesian information fusion with respect to a reference database (TSG-20).

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. Finally, we demonstrate that cross-modal top-down information processing – geo-information priming the recognition method in visual feature space – can yield even better results and more economic computing, verifying the advantage of using attentive processing in multimodal interfaces.

2 THE ATTENTIVE MACHINE INTERFACE

2.1 Related Work

In ubiquitous computing, several frameworks have been proposed in the frame of attentive interfaces and context awareness. [14, 1] proposed Attentive User Interfaces (AUI) that capture the attention of the user, e.g. from eye gaze estimation, and consequently adapt interaction systems for better communication with the user. [3] proposed that context is a description of a real world situation on an abstract level that is derived from available cues. [2] described the role of perceptual components in a context aware system for interaction. Finally, [11] 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 [11], it presents processing in a probabilistic framework and enables top-down, i.e., attentive cross-modal information processing.

Previous work on mobile vision services primarily advanced the state-of-the-art in computer vision methodology for the application in urban scenarios. [13] provided a first innovative attempt on building identification proposing local affine features for object match-

[1] JOANNEUM RESEARCH Forschungsgesellschaft mbH, Institute of Digital Image Processing, Wastiangasse 6, 8010 Graz, Austria, email: {katrin.amlacher,lucas.paletta}@joanneum.at

ing. [15] introduced image retrieval methodology for the indexing of visually relevant information from the web for mobile location recognition. Subsequent attempts [8, 10, 4] advanced the methodology further towards highly 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.

2.2 Concept and Architecture

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.

The architecture of the AMI reflects the enabling of both bottom-up and top-down information processing and would support snapshot (e.g., image) based as well as continuous operation on a stream of input data (e.g., video). 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 AMI. The user request for context information is handled by a Master Control (MC) component that schedules the processing (multiple users can start several tasks) and associates with each task corresponding system monitoring (SM) procedures. A concrete task is then performed by 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. Since such a subgraph can provide several ways of processing, the appropriate part can get selected by means of, e.g., time constraint, confidence of the expected result and quality of context-generating services. If a service gets offline, TP can switch to another similar service or to another processing chain, where already processed data is reused. The Context Graph Manager (CGM) maintains and manages context-generating modules in a graph structure (Context Graph). These 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. CGM assembles the subgraph according to several constraints, such as, task information, availability of context-generating modules and data and ensures that the subgraph is processable. The AMI functionality ensures the possibility to arbitrarily combine services and implements process flow regulation mechanism, e.g. when a service gets offline to switch to another service. It is also possible to invoke an additional processing chain if the confidence of the result it too low. Multiple users can concurrently request context information and the services are targeted towards fast and accurate (robust) responses.

2.3 Context Processing

For high-level context generation various services are required to combine information, services may temporarily exist, and outputs may be combined in arbitrary manner. The Context Graph – a directed acyclic graph with nodes representing individual *context processing* units, edges describing the information flow – is a flexible and extensible data structure that correspondingly connects between

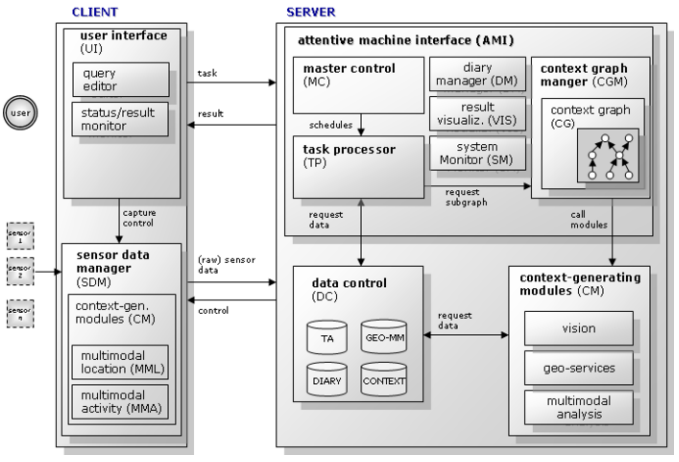


Figure 1. Concept of a client-server system architecture with attentive machine interface.

individual functionalities. Each context node provides a context-generating service that derives context information from its input data; context nodes are linked together depending on input and output data of the wrapped services; context nodes represent context information at a different level, where high-level context information is demanded by the user. 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. The resulting high-level context information is given to the user via a visualization compent and is stored in the Data Control or Diary Manager.

2.4 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 two different modes in performing the service of *geo-indexed object recognition* (Sec. 3). 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 and integration as outlined in more detail in Sec. 3.

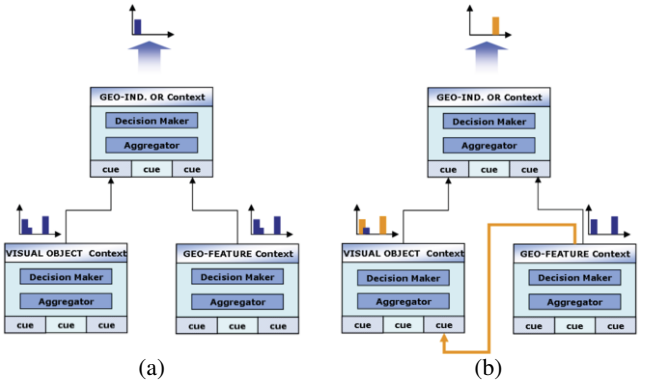


Figure 2. Two different processing modes visualised by their associated context subgraphs for "Geo-Indexed Building Recognition" (Sec. 3). (a) Bottom-up information processing of visual object recognition and geo-features. (b) Top-down information processing by using geo-features to prime visual object recognition (Sec. 2.4).

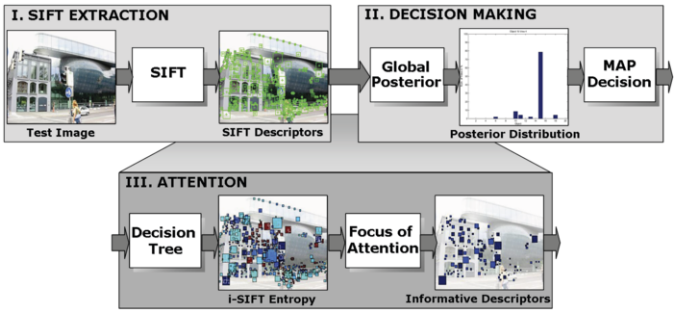


Figure 3. Concept for recognition from informative local descriptors. (I) SIFT descriptors are extracted within the test image. (II) Decision making analyzes the descriptor voting for MAP decision. (III) In i-SIFT attentive processing, a decision tree estimates the SIFT specific entropy; informative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

3 GEO-INDEXED OBJECT RECOGNITION

Urban 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 [5]. Cell identification is not treated here due to its large positioning error. Alternative concepts for localization are economically not affordable, such as, INS and markers that need to be massively distributed across the urban area. For image based urban object recognition, we briefly describe how we make use of the methodology presented in [12, 4]. 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 [9, 6] and the integration of local information into object recognition. The SIFT (Scale Invariant Feature Transformation) descriptor [6] is widely used for its capabilities for robust matching despite viewpoint, illumination and scale changes in the object image captures which is mandatory for mobile vision services. The *Informative Features Approach* (i-SIFT [4]) applied to mobile imagery in our experiments uses local density estimations to determine the posterior entropy, making local information content explicit with respect to object discrimination. The information content from a posterior distribution is determined 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 from the posterior distribution within a sample test point's local neighborhood in descriptor space. One is primarily interested to get the *information content* of any sample local descriptor $\mathbf{d}_i$ in descriptor space $\mathcal{D}$, $\mathbf{d}_i \in \mathcal{R}^{|\mathcal{D}|}$, with respect to the task of object recognition, where o_i denotes an object hypothesis from a given object set $\mathcal{S}_O$. For this, one needs 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 $H(O|\mathbf{d}_i) \equiv -\sum_k P(o_k|\mathbf{d}_i) \log P(o_k|\mathbf{d}_i)$. One approximates the posteriors at $\mathbf{d}_i$ using only samples $\mathbf{g}_j$ inside a Parzen window of a local neighborhood ϵ, $||\mathbf{d}_i - \mathbf{d}_j|| \leq \epsilon$, $j = 1 \ldots J$.

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, i.e., those $\mathbf{d}_i$ with $\hat{H}(O|\mathbf{d}_i) \leq H_\Theta$. For the rejection of images whenever they do not contain any objects of interest one considers 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.

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 [7], 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

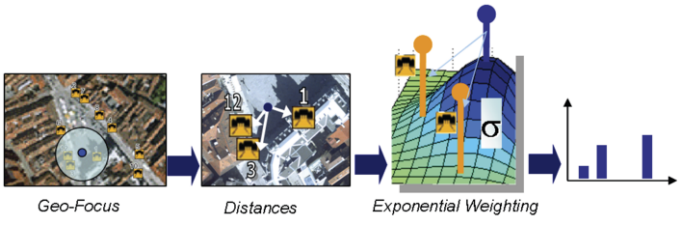


Figure 4. Extraction of object hypotheses from geo-services. (Left to right) Within a local spatial neighborhood (geo-focus), distances to the points of interest are determined, weighted by an exponential function and normalised to result in a distribution on object hypotheses.

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 receive 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 approaches in the AMI. 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 [4]) 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. In the experiments we used an ultra-mobile PC (Sony Vaio UMPC VGN-UX1XN) with 1.3 MPixels image captures. Reference imagery [4] with 640×480 resolution about building objects of the TSG-20 database[2] 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. Two additional views

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

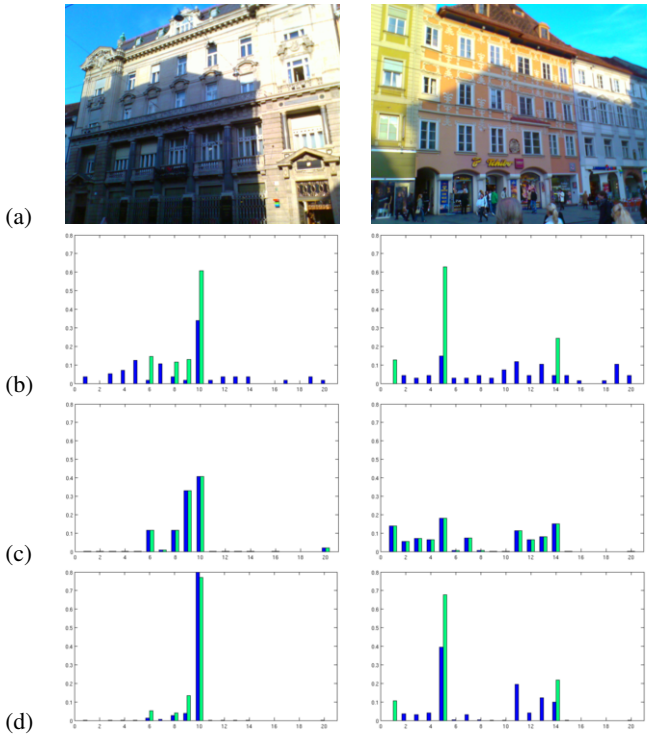
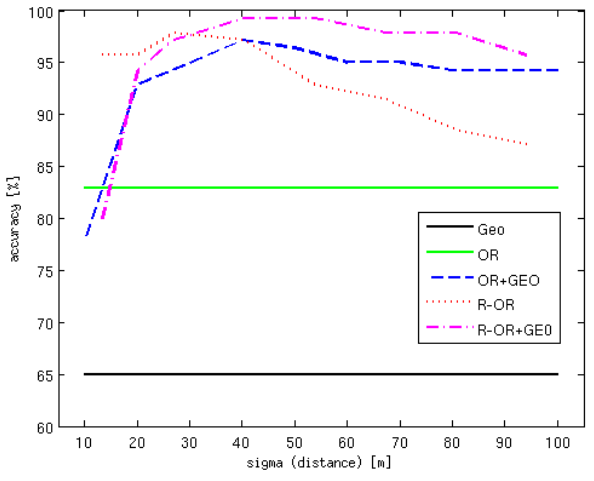


Figure 5. Comparison between bottom-up (blue/dark bars) and top-down approach (green/light bars) from (a) sample input images. Integration of object hypotheses from (b) vision and (c) geo-services into a (d) fused distribution demonstrates clear increases in the confidences of the correct object hypothesis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

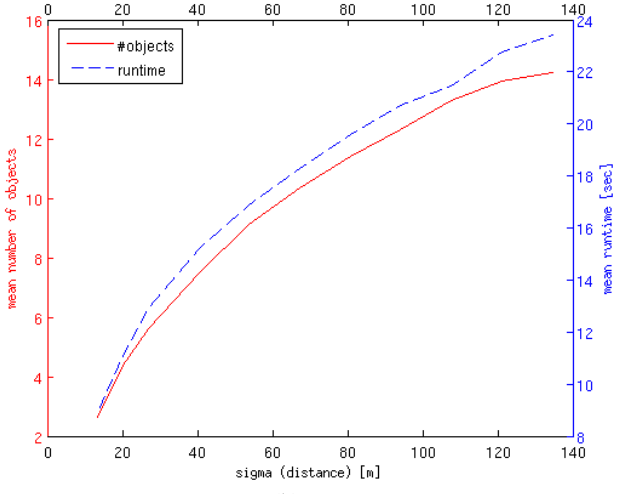
were taken for test purposes, giving 40 test images in total. For the evaluation of background detection we used a dataset of 120 query images, containing only images of buildings and street sides without TSG-20 objects. Another dataset was acquired with the UMPC, which consists of seven images per TSG-20 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 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



(a)



(b)

Figure 6. (a) 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). (b) Geo-indexed object recognition involves only a fraction of hypotheses and reduces computing time.

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 $\approx 96\%$. 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 improve 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 significant improvement in performance when using geo-services.

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

ACKNOWLEDGEMENTS

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 subproject S9104-N04.

REFERENCES

[1] Leonardo Bonanni, Chia-Hsun Lee, and Ted Selker, 'Attention-based design of augmented reality interfaces', in *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pp. 1228–1231, New York, NY, USA, (2005). ACM.

[2] James L. Crowley, Joëlle Coutaz, Gaeten Rey, and Patrick Reignier, 'Perceptual Components for Context Aware Computing', in *UBICOMP 2002, International Conference on Ubiquitous Computing*, Goteborg, Sweden, (September 2002).

[3] Anind K. Dey and Gregory 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).

[4] Gerald Fritz, Christin Seifert, and Lucas 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).

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

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

[7] 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*, pp. 61–64. Springer, (2005).

[8] Raphaël Marée, Pierre Geurts, Justus Piater, and Louis Wehenkel, 'Decision trees and random subwindows for object recognition', in *ICML workshop on Machine Learning Techniques for Processing Multimedia Content (MLMM2005)*, (2005).

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

[10] Stepan Obdrzalek and Jiri Matas, 'Sub-linear indexing for large scale object recognition.', in *Proceedings of the British Machine Vision Conference*, volume 1, pp. 1–10, (2005).

[11] Albrecht Schmidt and Kristof Van Laerhoven, 'How to build smart appliances', *IEEE Personal Communications*, 66 – 71, (2001).

[12] C. Seifert, G. Fritz, L. Paletta, and H. Bischof, 'Learning to focus attention on discriminative regions for object detection', in *Proc. European Conference on Artificial Intelligence, ECAI 2004*, pp. 932–936, (2004).

[13] 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*, pp. 71–80. Chicago,IL, (2003).

[14] Roel Vertegaal, 'Attentive User Interfaces', *Communications of the ACM*, **46**(3), 30–33, (2003).

[15] 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*, pp. 76–81, Washington, DC, (2004).