Scene Interpretation: Unified Modeling of Visual Context by Particle-Based Belief Propagation in Hierarchical Graphical Model

Sungho Kim and In So Kweon

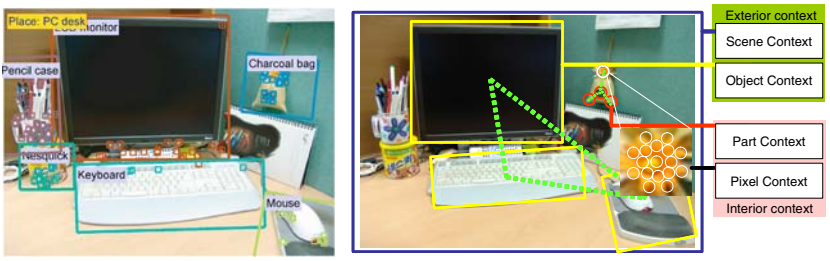
Dept. of EECS, Korea Advanced Institute of Science and Technology,
373-1 Gusong-Dong, Yuseong-Gu, Daejeon, Korea
{sunghokim, iskweon}@kaist.ac.kr

Abstract. In this paper, we present a novel scene interpretation method by unified modeling of visual context using a hierarchical graphical model. Scene interpretation through object recognition is difficult due to several sources of ambiguity (blur, clutter). We model the visual context of scene, object, and part to disambiguate them during recognition. A precisely designed hierarchical graphical model can represent the contexts in a unified way. We also propose a new inference method, particle-based belief propagation, optimized to scene interpretation in this hierarchical graphical model. Such an inference method suits the high-level context of scene interpretation. In addition, our core inference is so general that it can be used in any complex inference problems. Experimental results validate the power of the proposed model of visual context to solve the ambiguities in scene interpretation.

1 Introduction

The main task of scene interpretation in high level vision is to identify and determine the pose of 3D objects within a 2D image such as Fig. 1(a). A scene usually contains several types of 3D object in front of a complex background. The conventional local, feature-based object recognition methods [1][2][3], which use only individual object information, may work under high-quality viewing conditions, however, such methods often generate false alarms in ambiguous environments. In real, uncontrolled working environments, the ambiguities of scene interpretation originate from image blurring, background clutter and similarity of objects. Camera images can be blurred by short image acquisition time and large distances. Features from the background or other objects can cause false matching, which degrades object recognition performance. Previous works tried to remove the influence of background clutter by stereo matching-based figure-ground segmentation [4], distance ratio [1]. Another approach incorporates the background information rather than removing it. Torralba et al. propose a simple Bayesian formula using background features [5]. They get prior distribution of object label, position, and scale from background features. From the interpretation of many scene-images, we find a very interesting fact: many objects appear together and are strongly related to specific scenes.

P.J. Narayanan et al. (Eds.): ACCV 2006, LNCS 3852, pp. 963–972, 2006.



(a) (b)

Fig. 1. (a) Scene interpretation result of our system: Labeled part, object and place information is overlayed. (b) Four types of visual context such as scene, object, part, and pixel context are interrelated within a scene.

The relational information between scene and objects, and between objects, provides visual context in vision. Visual context can alleviate the recognition problem enormously. If we view only the separated objects in Fig. 2(a), we cannot discriminate between them because image blurring gives them similar shapes and appearances. However, if we view Fig. 2(b), we can recognize that the left object is a hair drier in a bathroom, and the right object is a drill in a workshop. Objects are usually defined by function and relation. Objects are associated with some scenes more than others, just as seagulls are associated with the sea. Although there are many kinds of visual context, we confine them to exterior context (scene,object context) and interior context (part, pixel context) as Fig. 1(b). According to cognitive experiments performed by Bar and Ullman [6], a spatial context between parts has substantial effect on recognition performance. Carbonetto proposed MRF-based modeling of spatial context in object layer only [7].

The key idea of this paper is to model this kind of relational information and use it to resolve ambiguities. Section 2 explains the details of the computational model of context in scene interpretation. Section 3 and 4 deal with an inference and a learning method respectively. Section 5 details the specific implementations. We validate the proposed method through large-scale experiment in Section 6 and conclude in Section 7.

(a) (b)

Fig. 2. (a)We cannot discriminate which one is a drier, which one is a drill without scene context. (b) We can discern them more accurately with the scene context [8].

2 Hierarchial Graphical Model of Visual Context

In this section, we present a novel framework, incorporating multiple visual contexts, to improve the efficiency and reliability of object recognition in ambiguous environments. Pixel context is used to build the visual features of local image patches. Spatial relations of each pixel's edge orientation, edge magnitude and color are encoded to form visual features. Part context prompts expectations for neighboring parts and objects. Object context provides expectations of neighboring objects and scene information such as place. Scene context provides the priors of object existence. These contexts interact with one another and exchange contextual information to provide reliable recognition results.

A graphical model is a suitable tool for dealing with such a complex system description. A graphical model is simply a marriage between probability theory and graph theory [9]. Nodes represent random variables, and the arcs or edges represent probabilistic interaction between variables. This can solve uncertainty and complexity problems simultaneously by compact representation of joint probability distribution. We have to estimate multiple variables, such as part identity and pose (x_P), object identity and pose (x_O), and scene properties like place identity (x_S). If we model this problem using a simple Bayesian framework with a simple directed graphical model, as Fig. 3(a), then we can represent the joint probability distribution in a factored form, as in equation (1) (This is conventional approach).

$$p(x_S, x_O, x_P, I) = p(I|x_P)p(x_P|x_O)p(x_O|x_S)p(x_S) \tag{1}$$

Although this graphical model can represent the joint probability density in a simpler form, it cannot model the whole visual context correctly. The first problem of the model is that it cannot represent the hierarchical interaction of each layer explicitly. Only top-down contexts are represented using directed arrows. However, in practice, bottom-up contextual information also exist. As indicated in [8], recognized objects can activate a scene context, and a recognized scene can also activate object recognition. Objects and object parts have properties of bidirectional exchange similar to the scene-objects case. The second problem is that neighboring contexts of parts and objects are not reflected in this graphical model. As Bar and Ullman showed when they demonstrated the importance of spatial relation in object recognition [6], we have to insert the spatial relation context or neighbor context in the part and object layer. Based on these cognitive facts, we solve the first problem by introducing an undirected graphical model, such as Markov Random Field (MRF), a generalized version of directed graphical model. MRF can more accurately represent the bidirectional property of each layer. We solve the second problem by adding more spatial nodes to reflect the neighboring context in the part and object layer.

Fig. 3(b) shows the refined graphical model for multiple context-based object recognition. This graphical model can represent all the contexts properly. Contexts are reflected on two types of graphical representations. The top-down and bottom-up context of hidden variables is handled in tree-structured graph-

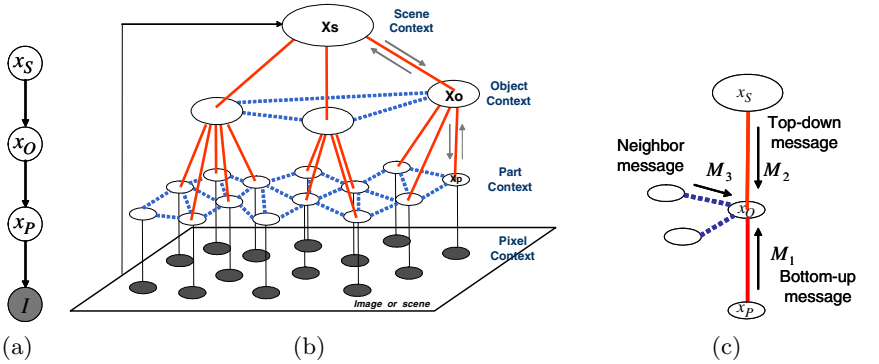


Fig. 3. (a) Simple Bayesian network can model only top-down influences. (b) Proposed hierarchical graphical model (HGM) model can represent bottom-up, top-down and neighboring context simultaneously. (c) An object node gathers three kinds of messages.

ical representation (here, red thick lines). In addition, sensory evidence is represented by thin black lines. The neighboring context of parts and object is reflected on planar loop structured graphic (here, dotted thick blue lines). The black nodes are pixel contexts acting as visual features robust to photometric and geometric distortions. These pixel contexts provide bottom-up evidence to the part layer. Similarly, whole scene features give bottom-up evidence to the scene layer.

3 Inference by Particle-Based Belief Propagation

3.1 Modified Belief Propagation (*BP*)

The goal of scene interpretation using the graphical model of Fig. 3(b) is to estimate hidden variables. We first assume discrete random variables for as part identity, object identity and scene identity. From a statistical view point, variable estimation is equivalent to computing certain marginal probabilities. The term *inference* means the computation of marginal probabilities. A practical inference method is belief propagation (BP), which is supposed to solve inference problem at least approximately [10]. We adapt the standard BP to the hierarchial graphical model in terms of three aspects.

(1) Function-based message categorization: We can represent the multiple contexts by three types of messages: bottom-up (M_1), top-down (M_2), and neighbor (M_3) messages. Fig. 3(c) shows a part of the graphical model in object layer. An object node receives messages from the lower node (part information), the higher node (scene context) and neighboring nodes (neighboring object) simultaneously. The belief at the object node is updated by

$$B(x_O) = \alpha M_1(x_O) M_2(x_O) M_3(x_O). \tag{2}$$

(2) Max-product rule: We use the max-product message update instead of sum-product in standard BP because the max-product shows a significantly better convergence [11].

(3) Approximation of message update: Message updating in standard BP is very inefficient since the node where message is propagated has to be excluded during message gathering and while other messages are recalculated. We make the message update efficient by replacing it with a current belief ($B(x_S)$) of that node:

$$M_2(x_O) \leftarrow \max_{x_S}\{\psi_{OS}(x_O, x_S)B(x_S)\} \tag{3}$$

where $\psi_{OS}(x_O, xS)$ is the compatibility or correlation function between two nodes. Contextual information is stored in this compatibility function. The message is propagated by tune-MAX. We tune all possibly transferable messages by multiplying current belief by the compatibility function, then only the maximal message is propagated to the node. The modified BP is held for both part layer and scene layer as object layer.

3.2 Particle-Based Belief Propagation (*PBP*)

In general, belief distribution of each node cannot be represented by parametric forms. A stochastic approximate inference must represent the distribution by a set of weighted samples. Conventionally, nonparametric BP is optimized to continuous random variables such as tracking or feature localizations [12]. We apply the concept of particle filter to the proposed HGM for object recogntion.

As discussed, there are many sources of ambiguities from object similarity, blurring by motion, and image noises. One solution to these ambiguities, in the computational approach, is not to jump to conclusions but to allow multiple high-probability values to stay available until longer feedbacks like visual context exert an influence. The concept of particle filtering is to compute a set of plausible guesses instead of a single guess to estimate a variable. These guesses are then assigned as weights to approximate a posterior distribution. Fig. 4(a) shows the particle-based BP in the object layer. A particle is composed of a hypothesized object ID and deterministically estimated object pose (scale, orientation, and position in image) relative to model CFCM . Each particle weight is updated by tune-max ($M_2(x_O^{(i)}) = \max_k\{\psi_{OS}(x_O^{(i)}, x_S^{(k)})B(x_S^{(k)})\}$). In general, a particle is generated using three kinds of correlation functions. After message update, particles are resampled using optimal resampling [13]. The same PBP also exists in the part layer, and the scene layer.

4 Learning of Compatibilities

The notion of learning in graphical model is the same as the learning of compatibility functions that relate two neighboring nodes. Fig. 4(b) shows seven compatibility functions to learn. Two evidence functions ($\phi(y, x_P), \phi(y, x_S)$), part-part compatibility ($\psi(x_P, x_P)$), part-object compatibility for bottom-up

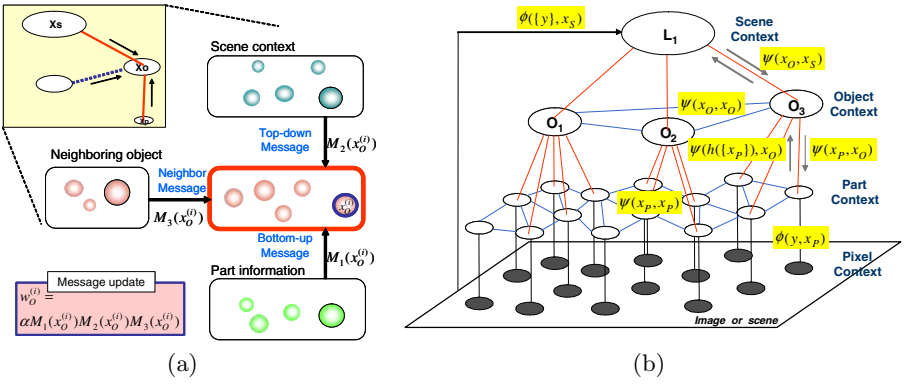


Fig. 4. (a) Each node is represented by a set of particles, or possible hypotheses. Belief of each particle is calculated by incoming bottom-up, top-down and neighboring messages. (b) Learning is estimating both nodes and compatibilities. There are 7 kinds of compatibilities to learn in the HGM.

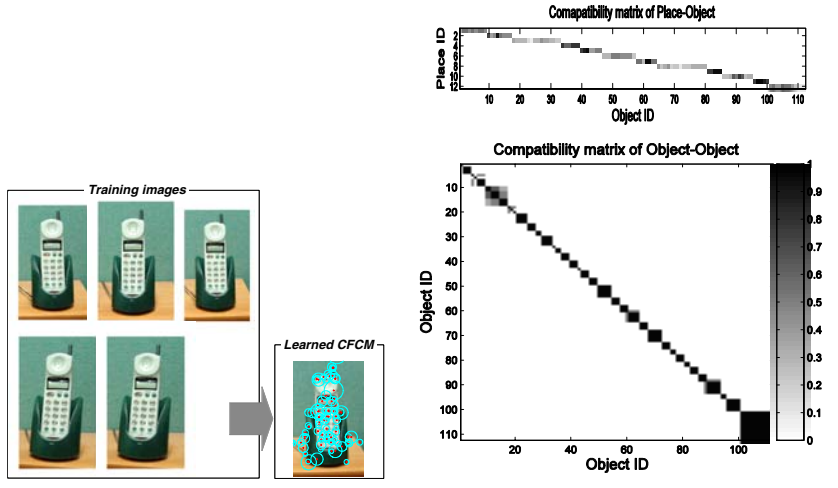
$(\psi(h(\{x_P\}), x_O))$, part-object compatibility for top-down $(\psi(x_P, x_O))$, object-object compatibility $(\psi(x_O, x_O))$, and scene-object compatibility $(\psi(x_O, x_S))$. These compatibilities can be regarded as functional representations of multiple visual contexts. The compatibility functions are modeled as follows:

- $\phi(y, x_P)$ is bottom-up evidence to part and estimated by Gaussian noisy measurement model of appearance similarity between scene and shared feature. Shared feature is generated by visual clustering in feature space.
- $\phi(\{y\}, x_S)$ is bottom-up evidence to scene and estimated by holistic voting of the distribution of nearest features. Each clustered scene feature contains the prior distribution of place.
- $\psi(x_P, x_P)$ is compatibility between neighboring parts and measured by same labeling and proximity of part location.
- $\psi(h(\{x_P\}), x_O)$ is compatibility between parts and object, which estimated through the size of Hough transform in pose space. Pose consistent parts provide messages in approximated form of Hough size.
- $\psi(x_P, x_O)$ is compatibility between part and object, which is estimated by modeling Gaussian noisy model of part pose.
- $\psi(x_O, x_O)$ is compatibility between objects and estimated by learning of labeled training objects.
- $\psi(x_O, x_S)$ is compatibility between object and scene. This is also estimated by counting labeled training images (see Fig. 5(b)).

5 Details of Implementation

5.1 Representation of Object and Scene

We interpret scenes at identification level: identifying previously viewed objects with place ID as in Fig. 1(a). We represent a 3D object with a set of view-



(a) (b)

Fig. 5. (a) An example of 3D object representation: 5 mutliview objects are clustered to a single CFCM. In a CFCM, each parts shares object pose parameters. (b) Compatibility matrices: (Top) shows place-object and (bottom) shows object-object compatibilities. Darker intensity represents stronger correlation.

clustered common frame constellation models (CFCM) that are extended to 3D object representation using [3][15] (see Fig. 5(a)). Each CFCM is composed of a set of learned parts. This means that each part contains both mean, variance of pose and an index to the shared features to handle a variety of objects. We assume that an object is decomposed into radial symmetry parts and corner-like parts. Features are generated by describing them with the localized histograms of edge orientation, edge density, and hue. This feature consists of a histogram vector of appearance and image structure-based pose (part size, part orientation, location) which is used to learn CFCMs. More details of the feature detector and scalable 3D object representation scheme are explained in [14] and [15], respectively. Place information is encoded into clustered features which store the distribution of place information.

5.2 Particle Management in Scene Interpretation

***Particle Generation*:** Ideally, we can generate particles using the compatibilities in bottom-up, top-down, and neighboring messages. However, we generate them using only bottom-up messages.

***Resampling particles*:** The recognition system degenerates to a single peak if we use unimodal particle representation. We solve this problem using multimodal particle representation in part layer and object layer [16].

***Final particle selection*:** The system requires at least four steps of concurrent message update and resampling to propagate the top-down context to the lowest

layer. Final scene interpretation is performed by selecting the max particles in each multi-modal representation.

6 Experimental Results

We evaluate the context-based scene interpretation system using a huge database. Table 1 summarizes the database. After scalable learning of 3D by feature clustering and view clustering, the feature size is reduced by 33.3% from 72,083 to 48,063 ($\varepsilon = 0.2$). After shared feature-based view clustering, the CFCM size is reduced from 5.5 CFCMs/object to 2.4 CFCMs/object ($T2$=10 pixels). Fig. 5 shows the learning results of compatibility between place-object and object-object by counting the occurrences.

The proposed system can remove the ambiguity of blurred object shown in Fig. 6. The place information acquired from overall scene features provides priors of certain objects. Finally, we evaluated our proposed method through extensive experiments with 228 indoor scenes. Recognition is assumed to be successful if both object ID and pose are correct. Fig. 7(a) is the results by cumulatively adding contexts. L1, L2, L3 represent part, object, scene layer, respectively. M1, M2, M3 represent bottom-up, top-down, neighboring message, respectively.

Table 1. Composition of database for training and test: We labeled place IDs to each images and objects are segmented and labeled for training. Test set is composed of unoverlapped images and unseened images (scene size: 640×480 color image).

Role		Scene		Object	
		No. of place	No. of scene	No. of objects	No. of views
Training		12	228 (even)	112	620
Test	Learned	12	228 (odd)	112	645
	Unlearned	random	25	0	0

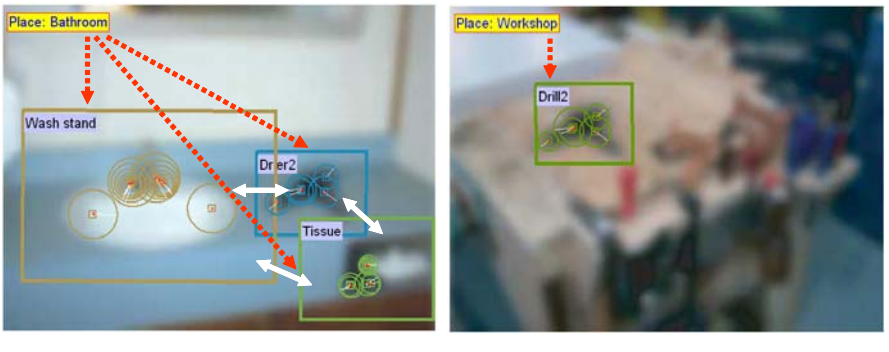


Fig. 6. The proposed context-based scene interpretation system can disambiguate blurred objects successfully, especially with the help of scene context

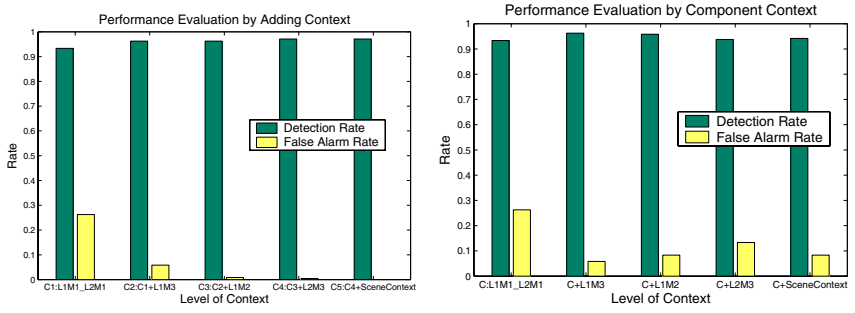


Fig. 7. (a) Performance by adding contexts: Full contexts show very low false alarm rate. (b) Component effect of individual context: Part context shows most dominant.

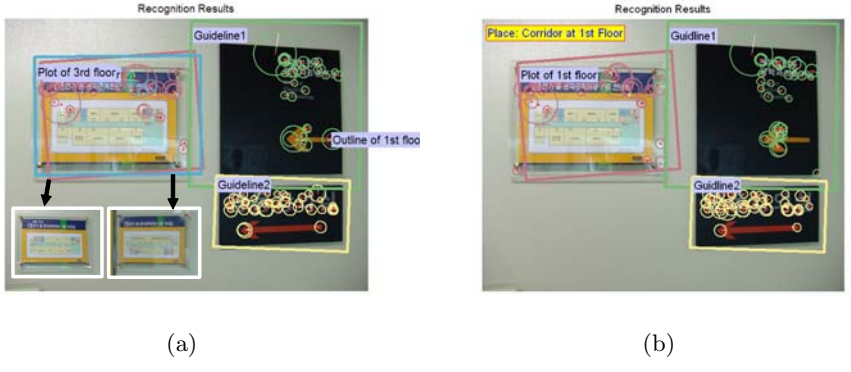


Fig. 8. Scene interpretation without scene context (a) and with scene context (b)

Especially C1 is basic recognition block which is composed of L1M1 and L2M1. So, L1M3 denotes neighboring part context, L1M2 denotestop-down context to part. L2M3 means neighboring object context. Without context, the detection rate (DR) is 95.8% and the false alarm rate (FAR) is 15%. However, if we use full context, the DR is 96.28% and FAR is 0.15%. Fig. 7(b) shows the impact of each context to recognition. Fig. 8 represents the power of scene context.

7 Conclusions

In this paper, we proposed a novel scene interpretation paradigm using the hierarchical context in cluttered indoor environments to remove ambiguities. The key contribution is unification of scene, object and part context using a hierarchical graphical model. To handle the ambiguities, we proposed a particle-based belief propagation method to object recognition problem. Finally, we validate the feasibility of model-based scene interpretation by the experiments in complex indoor environments. Work is underway to extend to the scene interpreta-

tion of category level by properly modeling feature detector and compatibility functions.

Acknowledgements

This research has been supported by the Korean Ministry of Science and Technology for National Research Laboratory Program (Grant number M1-0302-00-0064), Korea.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
2. Schmid, C.: A structured probabilistic model for recognition. In: CVPR'99, Fort Collins, Colorado, USA (1999) 485–490
3. Moreels, P., Maire, M., Perona, P.: Recognition by probabilistic hypothesis construction. In: ECCV '04. (2004) 55–68
4. Stein, A., Hebert, M.: Incorporating background invariance into feature-based object recognition. In: WACV'04. (2005) 37–44
5. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: ICCV'03, Washington, DC, USA (2003) 273–280
6. Bar, M., Ullman, S.: Spatial context in recognition. Perception **25** (1996) 324–352
7. Carbonetto, P., de Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. In: ECCV'04. (2004) 350–362
8. Bar, M.: Visual objects in context. Nature Reviews: Neuroscience **5** (2004) 617–629
9. Jordan, M.I.: Learning in Graphical Models. MIT Press (1999)
10. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Understanding belief propagation and its generalization. In G. Lakemayer and B. Nebel, editors, Exploring Artificial Intelligence in the New Milennium,Morgan Kauffmann (2002) 509–522
11. Weiss, Y., Freeman, W.T.: On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Trans. on Information Theory **47** (2001) 736–744
12. Sudderth, E.B., Ihler, A.T., Freeman, W.T., Willsky, A.S.: Nonparametric belief propagation. In: CVPR'03. (2003) 605–612
13. Fearnhead, P., Clifford, P.: On-line inference for hidden markov models via particle filters. J. R. Statist. Soc. B **65** (2003) 887–899
14. Kim, S., Kweon, I.S.: Biologically motivated perceptual feature: Generalized-robust invariant feature. In: ACCV'06. (2006) To appear
15. Kim, S., Kweon, I.S.: Scalable representation and learning for 3d object recognition using shared feature-based view clustering. In: ACCV'06. (2006) To appear
16. Vermaak, J., Doucet, A., Perez, P.: Maintaining multi-modality through mixture tracking. In: ICCV'03, Nice, France (2003)