10

An Effective 3D Target Recognition Imitating Robust Methods of the Human Visual System

Sungho Kim and In So Kweon
Korea Advanced Institute of Science and Technology
Korea

1. Introduction

Object recognition is an important research topic in computer vision. Not only it is the ultimate goal of computer vision, but is also useful to many applications, such as automatic target recognition (ATR), mobile robot localization, visual servoing, and guiding visually impaired people.

Great progress in this field has been made during the last 30 years. During 1970~1990, the research focused on the recognition of machine parts or polyhedral objects using edge or line information (Lowe, 2006, Faugeras & Hebert, 1986). A 2D invariant feature and hashing-based object recognition was popular during the 1990s (Mundy & Zisserman, 1992, Rothwell, 1993). Since the mid 1990s, view or appearance-based methods have become a popular approach in computer vision (Murase & Nayar, 1995). Current issues cover how to select a feature, handle occlusion, and cope with image variations in photometric and geometric distortions. Recently, object recognition methods based on a local visual patch showed successful performance in those environmental changes (Lowe, 2004, Rothganger et al., 2004, Fergus et al., 2003). But these approaches can work on textured complex object and do not provide 3D pose information of interesting objects.

The goal of our research is to get the identification and pose information of 3D objects or targets from either a visible or infrared band sensor in a cluttered environment. The conventional approaches as mentioned above do not provide satisfying results. To achieve this goal more effectively, we pay attention to the perception mechanism of the human visual system (HVS), which shows the best efficiency and robustness to the above mentioned problems. Especially, we focus on the components of HVS robustness.

2. Robust Properties of HVS

How have humans recognized objects robustly in a severe environment? What mechanisms cause a successful recognition of 3D objects? Based on these motivations, we researched various recent papers on psychophysical, physiological, and neuro-biological evidences and conclude the following facts:

Source: Vision Systems: Applications, ISBN 978-3-902613-01-1
Edited by: Goro Obinata and Ashish Dutta, pp. 608, I-Tech, Vienna, Austria, June 2007

2.1 Visual object representation in human brain

The HVS uses both view-based and model-based object representation (Peters, 2000). Initially, novel views of an object are memorized, and an object-centered model is generated through training many view-based representations. Another supporting evidence of this fact is that different visual tasks may require different types of representations. For identification, view-based representations are sufficient. 3D volume-based (or object centered) representations are especially useful for visual guidance of interactions with objects, like grasping them. In this paper, the goal is object identification and estimating the pose of objects for grabbing by a service robot. Therefore, both representations are suitable for our task.

2.2 Cooperative bottom-up and top-down information

Accordingly (Nichols & Newsome, 1999), not only the bottom-up process but also top-down information plays a crucial role in object recognition. Bottom-up process, called image-based, data-driven or discriminative process, begins with the visual information and analyses of smaller perception elements, then moves to higher levels. Top-down process is called knowledge-based perception, task dependent, or generative process. This process, such as high level context information (ex. place information) and expectation of the global shape, has an influence on object recognition (Siegel et al., 2000, Bar, 2004). So an image-based model is proper to the bottom-up and place context, and object-centered 3D model is suitable to top-down. The spatial attention is used to integrate separate feature maps in each process. From the detailed investigations in physiological and anatomical areas, many important functions of the bottom-up process were disclosed. Although the understanding of the neural mechanism of the top-down effects is still poor, it is certain that the object recognition is affected by both processes guided by the attention mechanism.

2.3 Robust visual feature extraction

(1) Hierarchical visual attention (Treisman, 1998): The HVS utilizes three kinds of hierarchical attention: spatial, feature and object. We utilize these attentions to the proposed system. Spatial attention is performed by a high curvature point like Harris corner, feature attention is made on local Zernike moments, and 3D object attention is done by the top-down process.

(2) Feature binding (Treisman, 1998): The binding problem concerns the way in which we select and integrate the separate features of objects in the correct combinations. Separate feature maps are bound by spatial visual attention. In the bottom-up process, we bind an edge map with a selected corner map and generate local structural parts. In the top-down process, we bind a gradient orientation map with gradient magnitude map focusing on a CAD model position.

(3) Contrast mechanism (VanRullen, 2003): Important information is not the amplitude of a visual signal, but is the contrast between this amplitude at a given point and at the surrounding location. This fact is true in the whole recognition process.

(4) Size-tuning process (Fiser et al., 2001): During object recognition, the visual system can tune in to an appropriate size sensitive to spatial extent, rather than to variations in spatial frequency. We use this concept for the automatic scale selection of the Harris corner.

(5) Part-based representation (Biederman, 1987): Visual perception can be done from part information supported by RBC (recognition by components) theory. It is related to the

properties of V4 receptive field, where the convex part is used to represent visual information (Pasupathy & Connor, 2001). A part-based representation is very robust to occlusion and background clutter. We represent visual appearance by a set of robust visual part.

Motivated by these facts, many computational models were proposed in computer vision. Researchers of model-based vision regarded bottom-up/top-down processes as hypothesis/verification paradigms (Kuno et al., 1988, Zhu et al., 2000). To reduce computational complexity, visual attention mechanism is used (Milanese et al. 1994). Top-down constraint is used to recognize face and pose (Kumar, 2002). Recently, an interesting computational model (HMAX) was proposed based on the tuning and max operation of a simple cell and a complex cell, respectively (Serre & Riesenhuber, 2004). In a computer vision society, Tu et al. proposed a method of unifying segmentation, detection and recognition using boosting and prior information by learning (Tu et al., 2005). Although these approaches have their own advantages, they modeled only on partial evidences of human visual perception, and did not pay attention to the robust properties of HVS more closely.

In this paper, we propose a computationally plausible model of 3D object recognition, imitating the above properties of the HVS. Bottom-up and top-down information is processed by a visual attention mechanism and integrated under a statistical framework.

3. Graphical Model of 3D Object Recognition

3.1 Problem definition

A UAV (unmanned aerial vehicle) system, such as a guided missile, has to recognize an object ID (identity) and its pose from a single visible or infrared band sensor. The goal of this paper is to recognize target ID and its pose in a UAV system, using a forward-looking visible or infrared camera. The object pose information is necessary for precise targeting.

We want to find the object name (θ_{ID}), the object pose ($\theta_C : \theta_{yaw}, \theta_{pitch}, \theta_{roll}$) relative to camera coordinates in a 3D world, the object position ($\theta_P : \theta_x, \theta_y$) and the object scale ($\theta_D$) in a 2D image. This information is useful in various applications. Similar processes exist in a primary visual cortex: ventral stream (what pathway) and dorsal stream (where pathway). The recognition problem can be formulated as the Bayesian inference by

$$\begin{aligned} P(\boldsymbol{\theta} \mid I) &= P(\boldsymbol{\theta} \mid Z_L, Z_C) \propto P(Z_L \mid \boldsymbol{\theta}, Z_C) P(\boldsymbol{\theta} \mid Z_C) \\ &= P(Z_L \mid \theta_{ID}, \theta_C, \theta_D, \theta_P, Z_C) P(\theta_{ID}, \theta_C, \theta_D, \theta_P \mid Z_C) \\ &\mathit{where} \quad I = \{Z_L, Z_C\} \end{aligned} \qquad (1)$$

where $\boldsymbol{\theta}$ means the parameter set as explained, I denotes input image, and it is composed of two sets: Z_L for object related local features Z_C for place or scene related contextual features. The likelihood of the equation (1), the first factor $P(Z_L \mid \boldsymbol{\theta}, Z_C)$ represents the posterior distribution of local features, such as local structural patch, edge information given parameters and contextual information. There is a lot of contextual information, but we restrict the information as place context and a 3D global shape for our final goal. This information alleviates the search space and provides accurate pose information. The second factor $P(\boldsymbol{\theta} \mid Z_C)$ provides context-based priors on object ID, pose which are related to the scene information by learning. This can be represented as a graphical model in a general form as Figure 1 (Borgelt et al., 2001). Scene context information can be estimated in a discriminative way using contextual features Z_C . Using the prior learning between scene and objects, initial object probabilities can be obtained from sensor observation. Initial pose information is also estimated in a discriminative way. Given those initial parameters, fine pose tuning is performed using a 3D global shape and sensor measurements, such as gradient magnitude and gradient orientation.

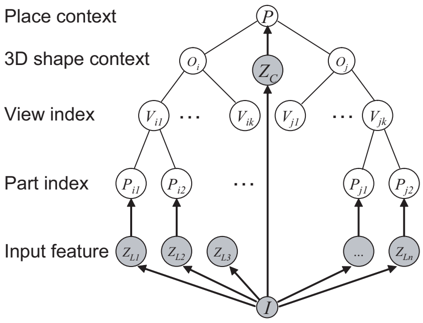


Figure 1. Graphical model of context-based object recognition: shaded circles mean observations and clear circles mean hidden variables

In the above graphical model, final parameters can be inferred from a discriminative method (bottom-up reasoning, such as directed arrows) and a generative method (top-down reasoning) with contextual information. To find an optimal solution from the equation (1), a MAP (maximum a posteriori) method is used generally. But it is difficult to obtain a correct posterior for a high dimensional parameter space (in our case 7 dimension). We bypass this problem by a statistical technique, drawing samples using a Markov Chain Monte Carlo (MCMC) technique (Green, 1996). The MCMC method is theoretically well-proved and a suitable global optimization tool for combining bottom-up and top-down information, which reveals superiority to genetic algorithm or simulated annealing although there are some analogies to the Monte Carlo method (Doucet et al., 2001). MCMC-like mechanism may not exist in the HVS, but it is a practically plausible inference technique in a high dimensional parameter space. Proposal samples generated from a bottom-up process achieve fast optimization or reduce burn-in time.

3.2 Basics of MCMC

A major problem of Bayesian inference is that obtaining the posterior distribution often requires the integration of high-dimensional functions. The Monte Carlo (or sampling) method approximates the posterior distribution as weighted particles or samples (Doucet et al., 2001, Ristic et al., 2004). The simplest kind is importance sampling, where random samples x are generated from $P(X)$, the prior distribution of hidden variables, and then weight the samples with their likelihood $P(y \mid x)$. A more efficient approach in high dimension is called the Markov Chain Monte Carlo (MCMC), a subset of particle filter. The Monte Carlo means samples and the Markov Chain means that the transition probability of samples depends only on a function of the most recent sample value. The theoretical advantage of the MCMC is that its samples are guaranteed to asymptotically approximate those which form the posterior. A particular implementation of the MCMC is the Metropolis-Hastings algorithm (Robert & Casella, 1999). The original algorithm is as follows:

Algorithm 1: Metropolis-Hastings algorithm

```
Draw an initial point $\theta_0$ from a starting distribution $P(\theta)$.
For i=1..N
        Draw candidate point $\theta_*$ from the jumping distribution $J_i(\theta_* \mid \theta_{i-1})$
        Calculate the ratio
                $\alpha = \frac{f(\theta_*) J_i(\theta_{i-1} \mid \theta_*)}{f(\theta_{i-1}) J_i(\theta_* \mid \theta_{i-1})}$
        Set $\theta_i = \theta_*$ with probability $\min(\alpha, 1)$, otherwise $\theta_i = \theta_{i-1}$
End for
```

The key concept of the algorithm is that the next sample is accepted with a probability of α. The next sample is obtained from jumping distribution or state transition function. Through the iteration, a sub-optimal solution can be obtained. However, the main problems of the method are a large burn-in time (the number of iterations until the chain approaches stationary) and poor mixing (staying in small regions of the parameter space for a long time). This can be overcome using domain information by the bottom-up process. Therefore, the finally modified algorithm is composed of the initialization part, calculated by the bottom-up process, and the optimization part obtained by the top-down process (see the Algorithm 2).

3.3 Object recognition structure

Figure 2 shows the proposed computational model of object recognition reflecting the robust properties of the HVS, as explained in section 2. Globally, bottom-up and top-down information is integrated under the statistical framework, MCMC. The object is represented as appearance-based in bottom-up, and object-centered in top-down. Furthermore, these object models are related to the scene context. Spatial attention is used to combine low-level feature maps for both bottom-up (in a local structure feature extraction block) and top-down (in shape matching block) processes. Detail computational procedures of each block are explained in the next sections. (Alogrithm 2 will help you to understand the proposed method.)

From a computational viewpoint, the proposed MCMC consists of three components: initialization, MCMC sampling and optimization. The bottom-up process means accumulating evidence computed from local structures and discriminates scene identity. Based on the scene context and local structural information, initial parameters such as object ID, pose, position and scale are estimated. The initial parameters are used to activate the 3D shape context. The MCMC samples are generated by a jumping distribution, which represents state-transition probability. From this sample, a 3D shape model is rendered. The final decision of object recognition is made after iterative sample generation and global shape matching. The decision information is fed back to the bottom-up process for another object recognition in the same scene. Algorithm 2 summarizes the overall recognition steps.

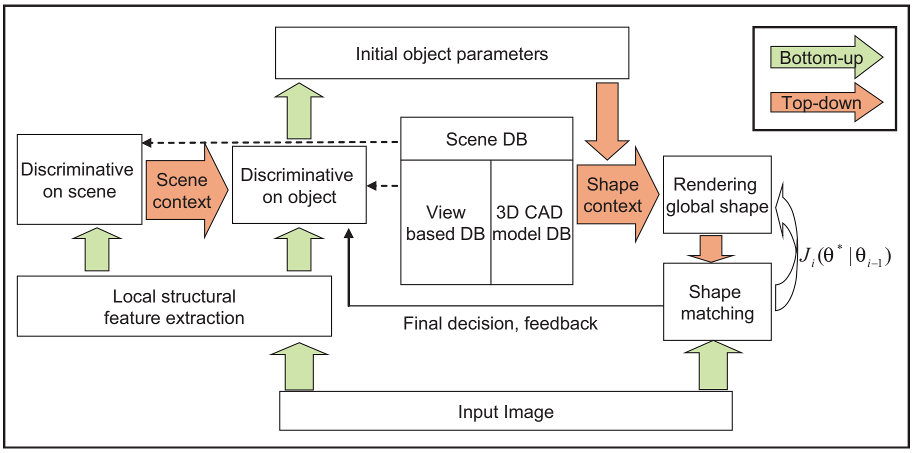


Figure 2. Overall functional model of the object recognition motivated by the robust properties of the HVS

Algorithm 2: Domain knowledge & context-based 3D object recognition algorithm

```
Stage I: Initialization by bottom-up process
            Step 1: Extract HCM, CEM in scale space
            Step 2: Find salient interesting points through scale space analysis.
            Step 3: Bind feature maps by relating salient HCM and the corresponding CEM
            Step 4: Extract local edge patches and calculate local Zernike moments
            Step 5: Discriminate scene ID through direct voting
            Step 6: Calculate the likelihood of object parameters from scene context and object
            discrimination by direct voting
            Step 6: Sort candidate parameters $\boldsymbol{\theta}_0 = \{\theta_{ID}{}^0, \theta_C{}^0, \theta_P{}^0, \theta_D{}^0\}$
Stage II: Optimization by top-down process
            Step 1: Extract GMM and GOM
            Step 2: Set initial point $\boldsymbol{\theta}_0 = \{\theta_{ID}{}^0, \theta_C{}^0, \theta_P{}^0, \theta_D{}^0\}$ from Stage I
            Step 3: Optimize parameters by MCMC sampling with feature map binding
                     For t = 0, …, T
                        Draw a candidate point $\boldsymbol{\theta}_*$ from the jumping distribution $J_t(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{t-1})$
                     Render the 3D CAD model based on shape context and $\boldsymbol{\theta}_*$
               Calculate the cost function $f(\boldsymbol{\theta}_*)$, by focusing on the rendered model
               and the integrated feature maps (GMM+GOM)
                Calculate the ratio
                                 $r = \frac{f(\boldsymbol{\theta}_*) J_t(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}_*)}{f(\boldsymbol{\theta}_{t-1}) J_t(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{t-1})}$
                        Accept $\boldsymbol{\theta}_t = \boldsymbol{\theta}_*$ with probability min(r, 1), or $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$
```

```
        End for
    Step 4: If $f(\mathbf{\theta}_T) < \varepsilon$, recognition finished and fed back to the step 6 in Stage I.
        Else reject $\mathbf{\theta}_0$ and go to step 2 with the next candidate $\mathbf{\theta}_0$
```

4. Scene Context-based Database

Figure 3 shows the scene-context-based database which is composed of object-specific scenes, 3D object models and view-based visual parts and their corresponding graphical model. It is displayed on the left.

4.1 Scene database

Conventional object recognition methods usually tried to remove background information. However, the background information of a scene provides important cues to the existence of target objects which are static or immovable, such as buildings and bridges. We call this information scene context. Learning the scene context is simple. First, we store various scenes which contain an interesting object. Then local visual features are extracted and clustered. (Details are explained in the next section.) Finally, clustered features are labeled with a specific object name and stored in a database. This database is used to recognize scenes as in Figure 2.

4.2 Object-centered model representation

As we discussed in section 2, the HVS memorizes object models in an object-centered way through enormous training. A plausible computational model is a 3D CAD model constructed manually. In this paper, we use a simple wireframe model for global shape representation. This method is suitable for man-made rigid objects like buildings, bridges, and etc. A voxel-based 3D representation may be appropriate for a generally shaped 3D object. The global 3D shape model provides the information of shape context which is useful to get the pose information and decision of the existence in the top-down process as Figure 2.

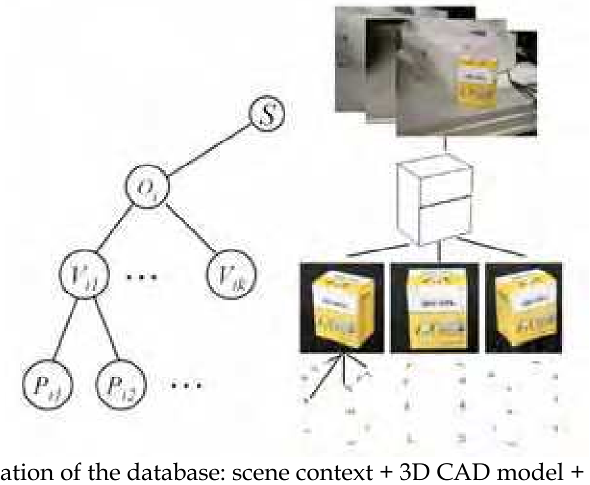


Figure 3. Configuration of the database: scene context + 3D CAD model + part-based view representation

4.3 View-based model representation

Basically, the HVS memorizes objects in an orientation dependent, view-based or appearance-based way (Edelman & Bülthoff, 1992). We quantize the view sphere by 30° and store each view as in Figure 3. Then, local visual parts for each view are extracted and represented using the proposed local feature. (Details will be explained in the next section).

5. Initialization by Bottom-up Process

A functional computational bottom-up process can be modeled as shown in Figure 2 (left half). Initial parameters are estimated through local feature extraction, discriminative method for scene recognition, and finally by discriminative process for object. Scene context provides prior information of a specific object ID which reduces the search space of the discriminative method for an object.

5.1 Local feature extraction

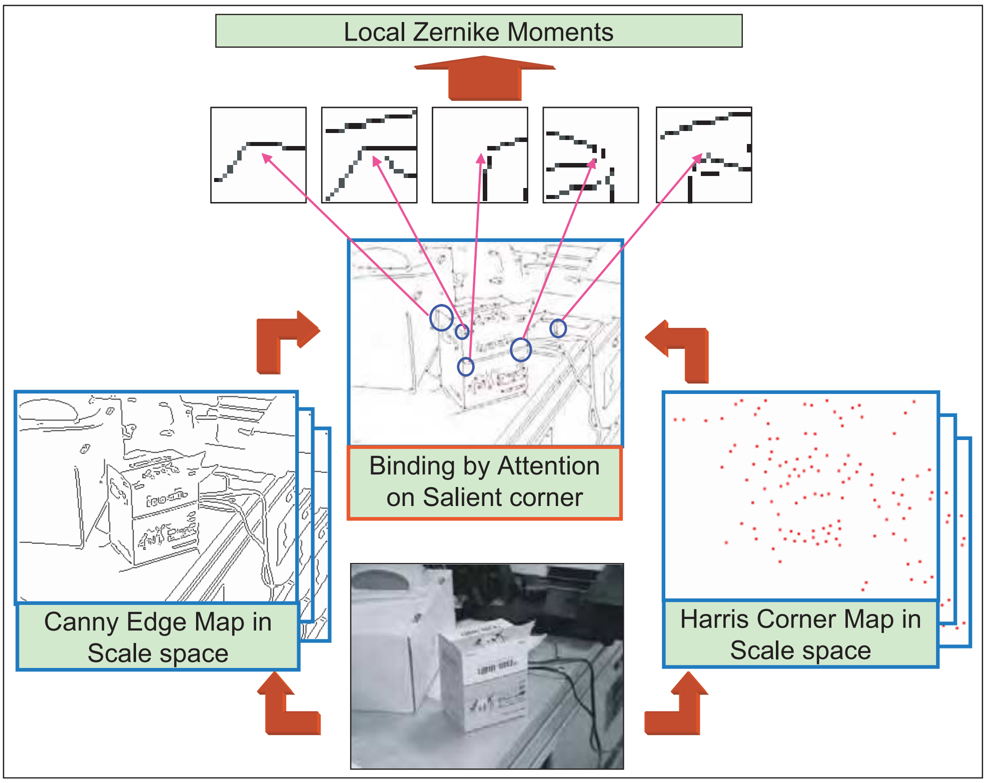

Local Zernike Moments
Binding by Attention on Salient corner
Canny Edge Map in Scale space
Harris Corner Map in Scale space

Figure 4. Block of local structural feature extraction: Canny Edge Map and Harris Corner Map are extracted in scale space which is bound by spatial attention on salient corners. Each local structural patch is represented using Zernike moments

Figure 4 shows the overall process for feature generation. We extract separate low-level feature maps such as Canny Edge Maps (called CEM) and Harris Corner Mapps (called HCM) in scale space. Then a perceptually salient corner and characteristic scale is calculated (Lindeberg, 1998). Locally structural visual parts are extracted by attending on CEM around salient corner points and scale tuned regions of HCM. The scale tuning process that exists is supported by the neuro-physiological evidence, as explained in section 2. Each patch whose size is normalized to 20×20 is represented by local Zernike moments introduced in (Kim & Kweon, 2005).

Step 1: Generation of separate feature maps

In the bottom-up process, we assume that an object is composed of local structures. According to (Parkhurst et al., 2002), Parkhurst et al. experimentally showed the fact that bottom-up saliency map-based attention of Itti's model is not suitable for learned object recognition. So, we adopt another spatial attention approach that the HVS usually attends on a high curvature point (Feldman & Singh, 2005). Although the HVS also attends on symmetrical points (Reisfeld et al., 1995), we only use the high curvature points for visual attention, since they are robust to a viewpoint and computationally easy to detect. We detect high curvature points directly from an intensity image using a scale-reflected Harris corner detector which shows highest repeatability in photometric, geometric distortions, and which contains enough information (Harris & Stephens, 1988, Schmid et al., 2000). A conventional Harris corner detector detects many clusters around a noisy and textured region. However, this doesn't matter, since the scale-reflected Harris detector extracts corners in noise removed images by Gaussian scale space. Furthermore, since salient corners are selected in scale space, corner clusters are rarely found, as in Figure 5. Canny edge detector is used to extract an edge map which reflects similar processing of a center-surround detection mechanism (Canny, 1986). The CEM is accurate and robust to noise. Both low level maps are extracted pre-attentively.

Step 2: Feature integration by attending on salient corners

Local visual parts are selected by giving spatial attention to a salient corner. We use the scale space maxima concept to detect salient corners. We define that a corner is salient if the measure of convexity (here, Laplacian) of corners in scale axis shows a local maxima. A computationally suitable algorithm is scale-adapted Harris-Laplace method which shows most robust to image variations (Schmid et al., 2000). Figure 5 shows the salient corner detection results. To detect a salient corner, first we make a corner scale space by changing the smoothing factor (σ). Then the convexity of corners are compared in scale axis.

Finally, salient corners are selected by selecting the maximum convexity measure in the tracked corners in scale space. As a by product, a scale tuned region can be obtained as Figure 5. This image patch corresponds to a local object structure.

Step 3: Local visual parts description by Zernike moments

The local visual parts are represented using modified Zernike moments introduced in (Kim & Kweon, 2005). The Zernike moments were used to represent characters because they are inherently rotation invariant, as well as possessing superior image representation properties, information redundancy, and noise characteristics. A normalized edge part is represented as 20 dimensional vectors where each element is the magnitude of a Zernike moment. Although we do not know how the HVS represents local visual image, we utilize the local Zernike moments, since this feature is robust to scale, rotation and illumination changes.

The performance is evaluated in terms of interest region selector and region descriptor using ROC curve (Mikolajczyk & Schmid, 2003). We used 20 object images as a reference, and made test images by changing the scale factor 0.8 times, planar rotation 45°, view-angle 25°, and illumination reduction by 0.7 time to the reference. For the comparison of the visual part detect, we used the same number of scale space, Zernike moment descriptor and image homography to check the correct matches. For the comparison of the descriptors, we use the same scale space, salient corner part detector and image homography for the same reason. Scale tuned region detector by the salient corner part detector almost outperform the SIFT (DoG-based) as in Figure 6 (a). In the descriptor comparison graph, SIFT and PCA show better performance than Zernike, as in Figure 6 (b). But this region of the low false positive rate is useless, because few features are found. In a noisy environment, our descriptor (Zernike) shows better performance. Figure 7 shows several matching examples using the salient corner with Zernike moments. Note the robust matching results in various environments.

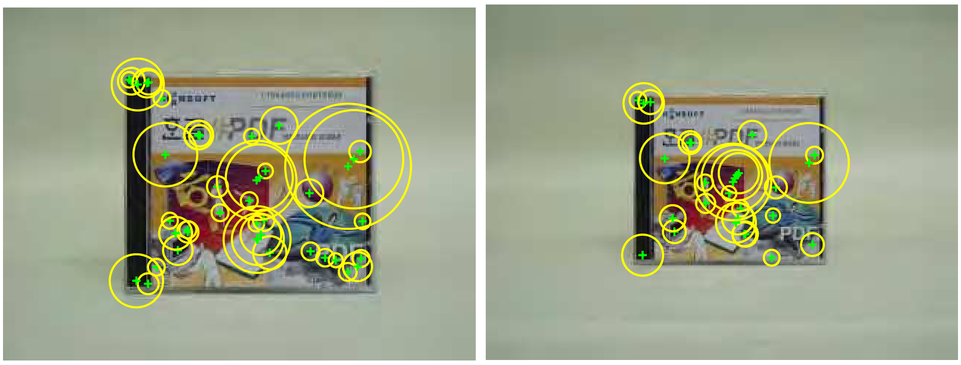
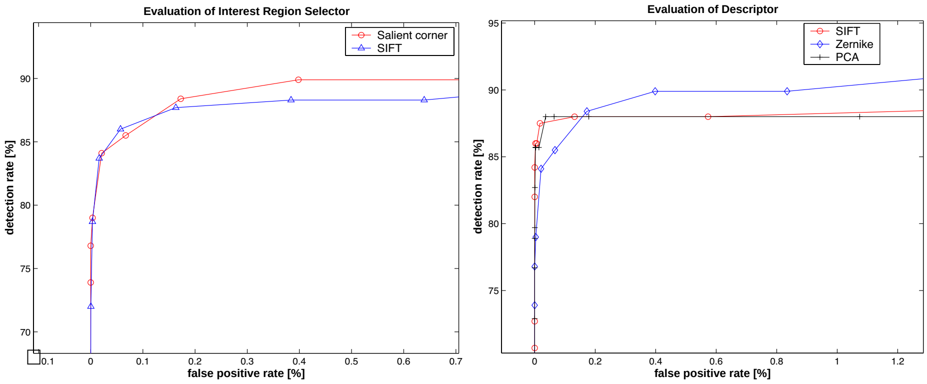


Figure 5. Examples of salient corners on a different scale



(a) (b)

Figure 6. (a) Performance comparison of interest part selector: Salient corner vs. SIFT, (b) performance comparison of local descriptor: SIFT, Zernike, and PCA

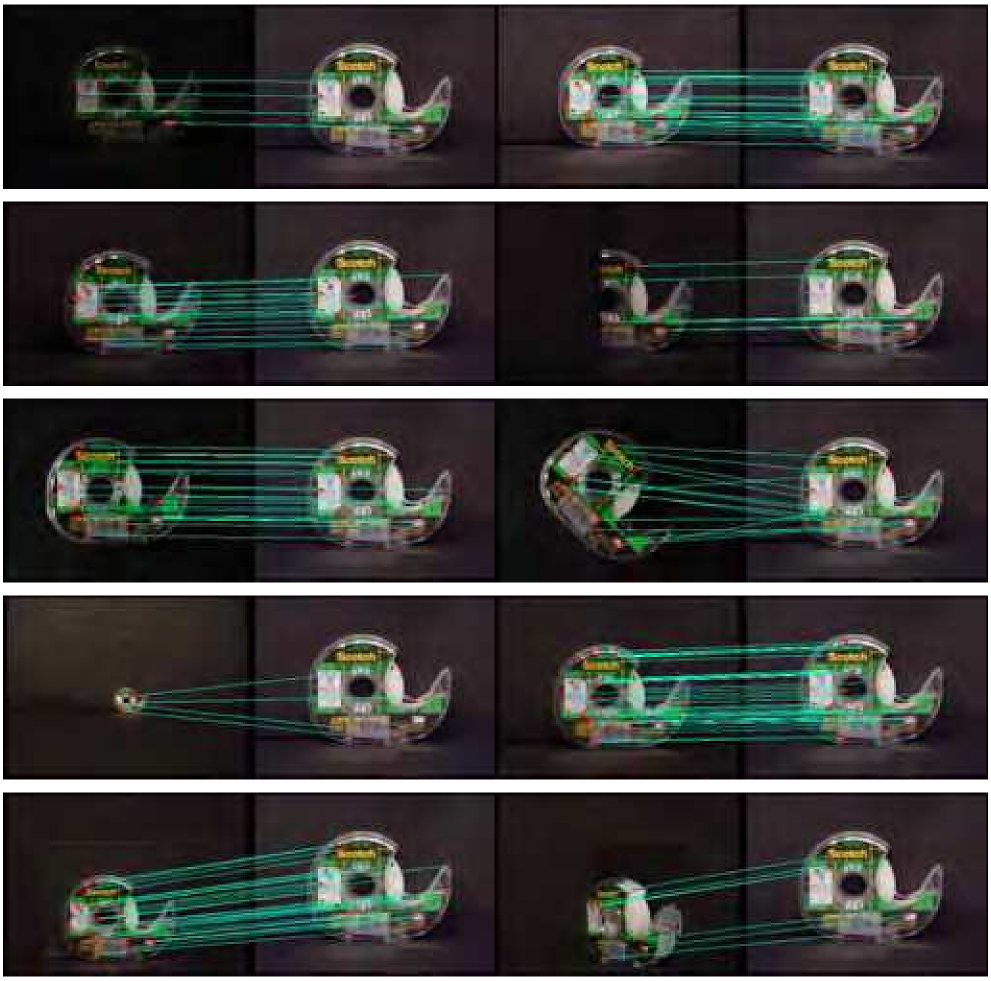


Figure 7. Examples of feature matching using a salient corner part detector and a Zernike moments descriptor in illumination, occlusion, rotation, scale and view angle changes

5.2 Initial parameter estimation by discriminative method

The initial parameters of an object are estimated using a discriminative method, 1-nearest neighbor based voting. In the first step, scene identity is found using direct voting. This scene context provides the information of probable object ID. In the next step, other initial pose, position, and scale parameters are estimated for the object, using the same voting method.

Step 1: Discriminative method on scene recognition,

In equation (1), the scene context term $P(\boldsymbol{\theta} \mid Z_C)$ provides object related priors especially object ID. If we assume one object per scene for simplification, then initial object ID can be estimated directly from the scene discrimination process as equation (2).

$$P(\theta_{ID} \mid Z_C) \approx P(s \mid Z_C) \tag{2}$$

The scene discrimination can be modeled as follows:

$$\theta_{ID} \sim S = \arg\max_{s} P(s \mid Z_C) \approx \arg\max_{l} \sum_{i=1}^{N_{Z_C}} P(s \mid Z_C^i) \tag{3}$$

where local feature $\mathrm{Z}_\mathrm{C}^\mathrm{i}$ belongs to scene feature set Z_C, which usually corresponds to background features. s is a scene label and N_{Z_C} is the number of input scene features. The posterior $P(s \mid Z_C)$ is approximated by the sum rule. We use the following binary probability model to design $P(s \mid Z_C^i)$:

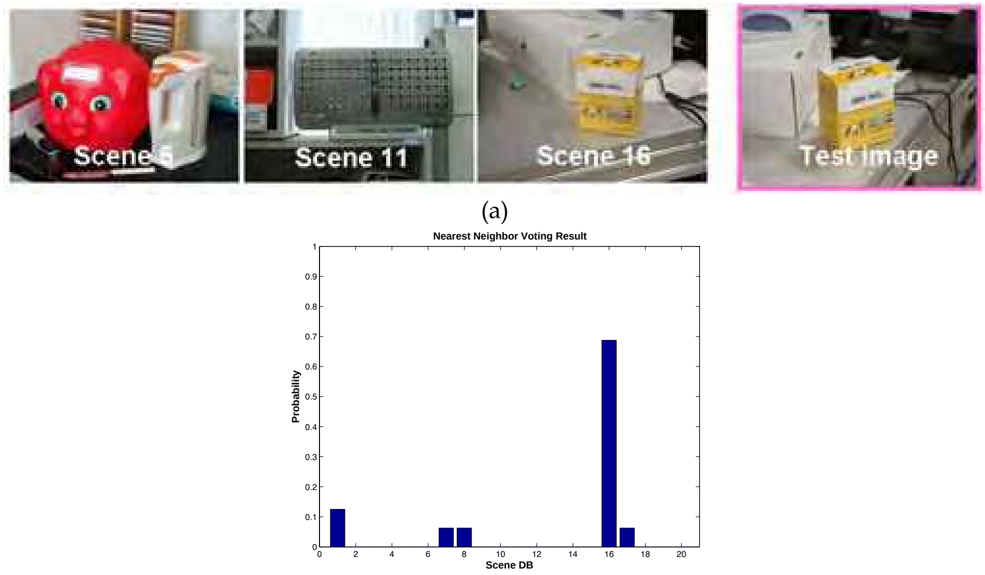
$$P(s \mid Z_C^i) = \begin{cases} 1 & L(Z_C^i) \in s,\ K_E(Z_C^i, \hat{Z}^i) \geq \delta \\ 0 & \mathit{otherwise} \end{cases} \tag{4}$$

where $L(Z_C^i)$ denotes the label of feature Z_C^i searched by 1-nearest neighbor search and $K_E(Z_C^i, \hat{Z}^i)$ is Gaussian Kernel of Euclidean distance between input feature Z_C^i and corresponding scene DB feature $\hat{Z}^i$. The kernel threshold δ usually set to 0.7~0.8. The final scene discrimination result provides scene context, prior information of object ID.

Step 2: Discriminative method on initial object parameters

Initial object ID is directly estimated from the scene context as step 1. Other object-related parameters are estimated by the same voting on view-based object DB. In equation (1), the initial parameters used in $P(Z_L \mid \theta, Z_C)$ can be directly discriminated as step 1, the voting scheme. Since we already know the initial object ID, the search space of other parameters are reduced enormously. The only difference is that the voting spaces are dependent on the parameters. For example, if we want to estimate the initial pose θ_C, we vote the nearest match pairs to the corresponding pose space (azimuth, elevation) like equation (3), and select the max. Given the initial object ID and pose, the initial object scale θ_D, and position θ_P is estimated easily, since our part detectors extract characteristic part scale with its position in the image (see Figure 5). So, the initial scale is just the average of the characteristic scale ratio between scene and model image, and the initial object position is the mean of matching feature pairs (see Figure 5). Since object parameters are estimated based on salient feature and scene context which reduce the search space, there is no increase of estimation error. Figure 8 shows the sample scene database and scene discrimination result by direct voting for the test image. In this test, we used 20 scenes in canonical view points for database and the test image was captured on a different view point. The scene 16 is selected by max operation of the voting result. This scene contains the interesting object. So, we can initialize the object ID parameter from this scene context.

Figure 9 shows a bottom-up result, where the 3D CAD model is overlaid using the initial parameters. There are some pose, scale, location errors. In addition, we cannot trust the estimated object ID. These ambiguities are solved through a top-down process using 3D shape context information.



(a)

(b)

Figure 8. (a) Examples of scene DB and test image on the right, (b) Scene context: nearest neighbor-based direct voting

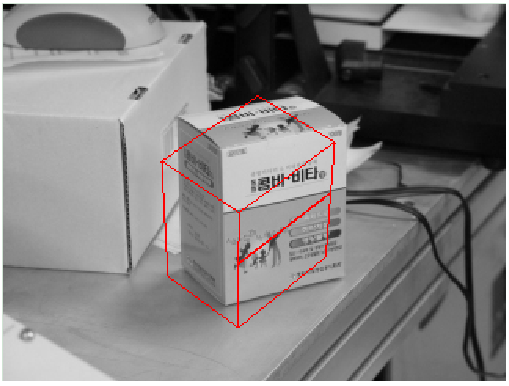


Figure 9. Initially estimated parameters by a bottom-up process

6. Optimization and Verification by Top-down Process

The Top-down process is crucial in the HVS. Although some top-down knowledge such as scene context information was already used for object discrimination, other context information like the expectation of a global 3D shape also has an important role in achieving more precise and accurate recognition. Figure 10 (or Figure 2: half right) shows the functional top-down procedures based on shape context initiated by a bottom-up process. Main components are model parameter prediction by jumping distribution and a global 2D shape matched by attending a shape model to combine gradient magnitude map (GMM) and gradient orientation map (GOM). The model parameter prediction and shape matching are processed iteratively for statistical optimization.

6.1 Generation of model parameters

A posteriori in equation (1) is approximated statistically by MCMC sampling. Based on the initial parameters obtained in bottom-up process, the next samples are generated based on the jumping distribution, $J_i(\theta_i \mid \theta_{i-1})$. It is referred to as proposal or candidate-generation function for its role. Generally, random samples are generated to prevent local maxima. However, we utilize the bottom-up information and top-down verification result for suitable sample generation. In this paper, we use three kinds of jumping types, i.e., object addition, deletion and refinement as Table 1.

The first type is to insert a new object and its parameters, depending on the result of a bottom-up process. The second is to remove a tested model and its parameters, determined by the result of top-down recognition. A jumping example of the third type is like equation (5). Next state depends on current state and random gain. This gain has uniform distribution (U) in the range of 30°, because the view sphere is quantized with this range. Here, θ_C^0 is initialized by the result of a bottom-up process.

$$\theta_C^t = \theta_C^{t-1} + \Delta\theta_C \tag{5}$$

$$\text{where } \theta_C = \begin{bmatrix} \theta_{yaw} & \theta_{pitch} & \theta_{roll} \end{bmatrix}^T, \Delta\theta_C \sim U(-15,15).$$

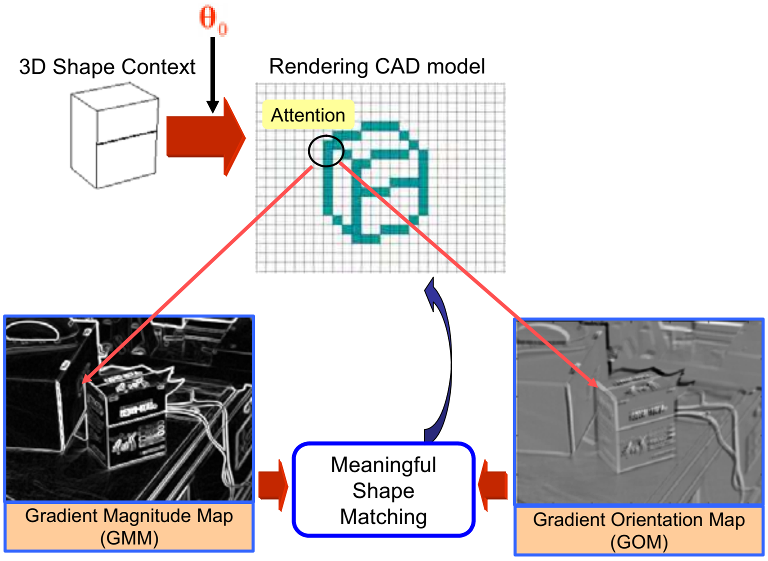

θ_0
3D Shape Context
Rendering CAD model
Attention
Meaningful Shape Matching
Gradient Magnitude Map (GMM)
Gradient Orientation Map (GOM)

Figure 10. 3D shape context-based top-down shape matching using MCMC: the 3D CAD model is rendered using the initial object parameters, then meaningful shape matching is performed by attending on the rendered 2D shape location and GMM, GOM. Final decision is made based on the MCMC optimization value

Jump type	Function	Parameters	Jumping distribution
J1	Object addition	$\theta_{ID}, \theta_C, \theta_D, \theta_P$,	Depend on bottom-up information
J2	Object deletion	$\theta_{ID}, \theta_C, \theta_D, \theta_P$	Depend on top-down result
J3	Fine tuning of parameters	$d\theta_C, d\theta_D, d\theta_P$	$\theta_C = \{\theta_{yaw}, \theta_{pitch}, \theta_{roll}\}$ $d\theta_{yaw} \in U(-30,30)$ $d\theta_{pitch} \in U(-30,30)$ $d\theta_{roll} \in U(-10,10)$ $d\theta_D \in U(\theta_D - \theta_D/5, \theta_D + \theta_D/5)$ $\theta_P = \{\theta_x, \theta_y\}$ $d\theta_x = U\{-40,40\}$ $d\theta_y = U\{-40,40\}$

Table 1. Jumping types and corresponding distributions

6.2 Robust shape matching

A predicted 3D CAD model generated by jumping distribution is rendered on the GMM and GOM image. Attending on the shape model points, both map information is combined as Figure 10. The scoring function used in the MCMC algorithm is defined by the shape matching. The shape matching between the rendered 2D shape and both maps is based on the computational gestalt theory (Desolneux et al., 2004). We propose a novel ε -meaningful shape matching method motivated from this theory.

Two important concept of the theory is as follows:

- Helmholtz principle: This principle provides a suitable mathematical tool for modeling computational Gestalt. Basically, it assumes that an image has random distribution of pixel values or orientations. If some pixels break the randomness, then these pixels have a certain pattern, called gestalt.
- ε -meaningful event: A certain configuration is ε -meaningful if the expectation in an image of the number of occurrences of the event is less than ε .

ε -meaningful shape matching

Since we only deal with intensity image or infrared image, all the available local information is just these three.

- Pixel intensity: $u(x,y)$
- Gradient magnitude: $u'(x,y) = \left(\frac{\partial u}{\partial x}, \ \frac{\partial u}{\partial y}\right)(x,y)$
- Gradient orientation: $\theta(x,y) = \frac{1}{\|Du(x,y)\|}\left(-\frac{\partial u}{\partial y}, \ \frac{\partial u}{\partial x}\right)(x,y)$

The last two components are useful for shape matching, since they are robust to illumination and noise. If we assume the image is random, then we can measure the structural alignment to a certain pattern. We can think of a matching at x_i that satisfy both the image gradient and orientation. If the rendered shape model is compatible to the image gradient and orientation simultaneously, then this matching is meaningful.

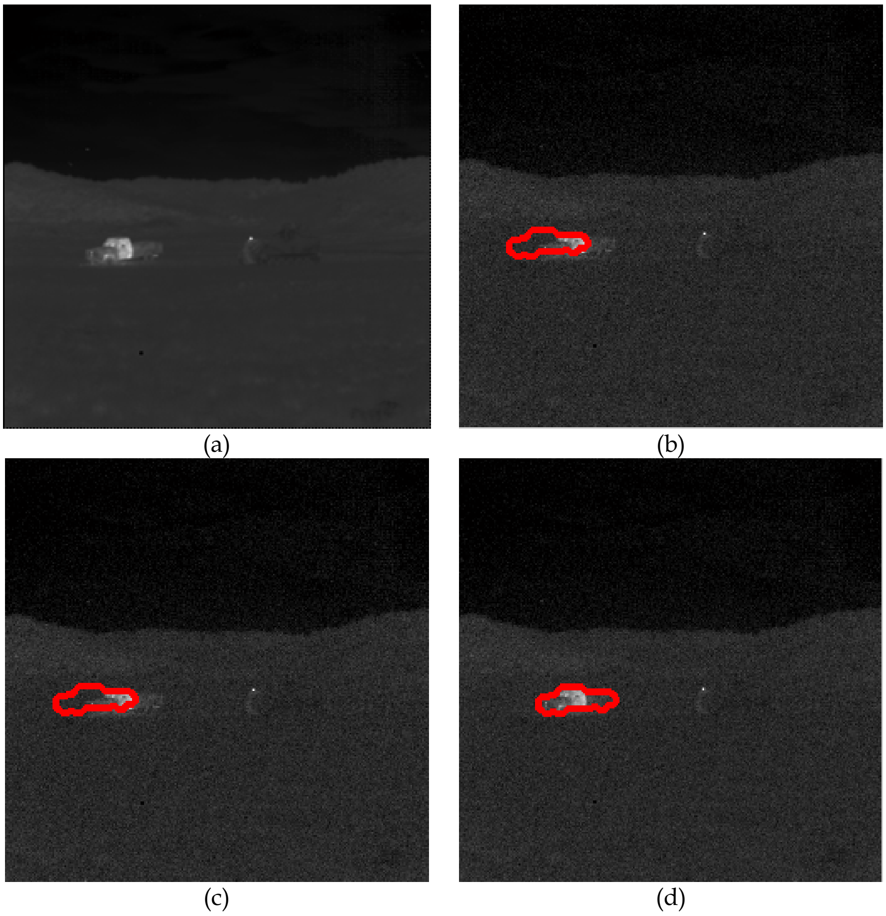


Figure 11. Shape matching examples on (http://www.cs.colostate.edu/~vision/ft_carson/): (a) original FLIR image, (b) GMM only, (c) GOM only, (d) proposed GMM+GOM

If the length of the rendered 2D shape is l, the probability of the event that gradient values ($C(\mathrm{x})$) are larger than a certain value, and orientation differences ($O(\mathrm{x})$) are within a precision along the shape model is defined in equation (6). The orientation precision is set to 8 directions.

$$P\left[C(\mathrm{x}_1)\ge\mu, O(\mathrm{x}_1)\le\frac{\pi}{4}\right]\cdot P\left[C(\mathrm{x}_2)\ge\mu, O(\mathrm{x}_2)\le\frac{\pi}{4}\right]\cdots P\left[C(\mathrm{x}_l)\ge\mu, O(\mathrm{x}_l)\le\frac{\pi}{4}\right]=H(\mathrm{x},\mu)^l \qquad (6)$$

where, $H(\mathrm{x},\mu)=\frac{1}{8}\cdot\frac{\text{num of }\{\mathrm{x}\mid C(\mathrm{x})\ge\mu\}}{\text{total image size}}$,

$C(\mathrm{x})=\|u'_I(x)\|$, $O(x)=|\theta_I(\mathrm{x})-\theta_M(\mathrm{x})|$, I for input, M for model,x:(x,y).

Definition: We call a matching between an image and a certain model is ε -meaningful shape matching if

$$f(\boldsymbol{\theta}) = N \times H(\mathrm{x}, \mu)^{l} \le \varepsilon \qquad (7)$$

where N is the number of the test. The smaller this value is, the better the shape matching is. We use this ε -meaningful shape matching as a scoring function for the MCMC optimization method because this function provides a measure of shape matching. The Scoring or cost function acts as a means of measuring the goodness of the proposed model parameters. Generated samples are accepted or rejected based on this function.

Figure11 shows the effectiveness of feature map binding in a top-down process. To show the power of feature map binding, we added Gaussian noise with a standard deviation 8. The binding GMM with GOM outperforms the single map based shape matching.

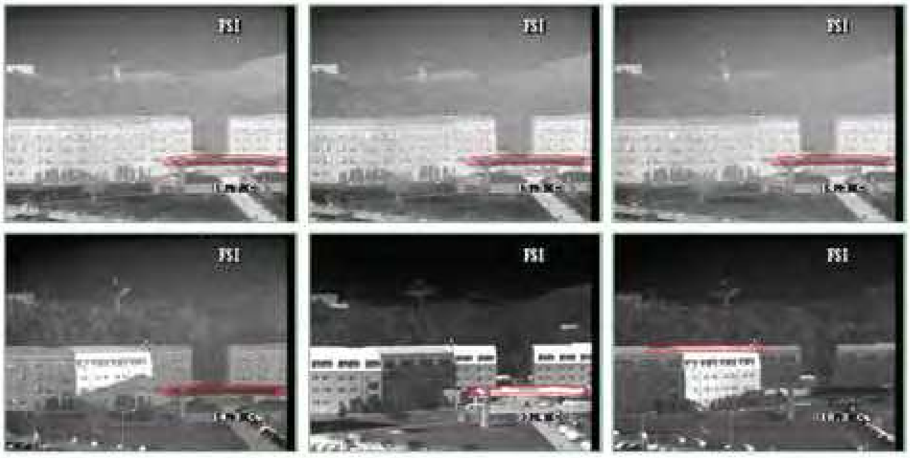
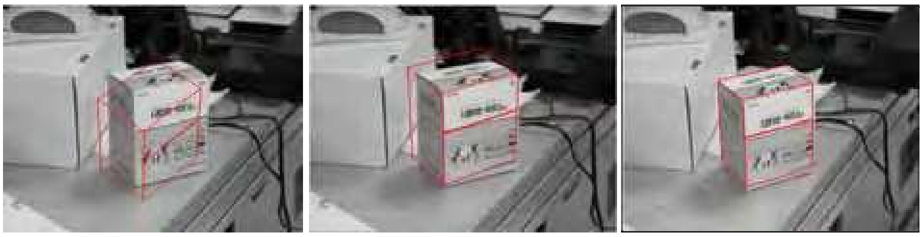


Figure 12. Shape matching results for temperature varying FLIR sequences. The proposed method is very robust to temperature changes. The last image shows a false matching result where the roof target hardly detectable by human eyes



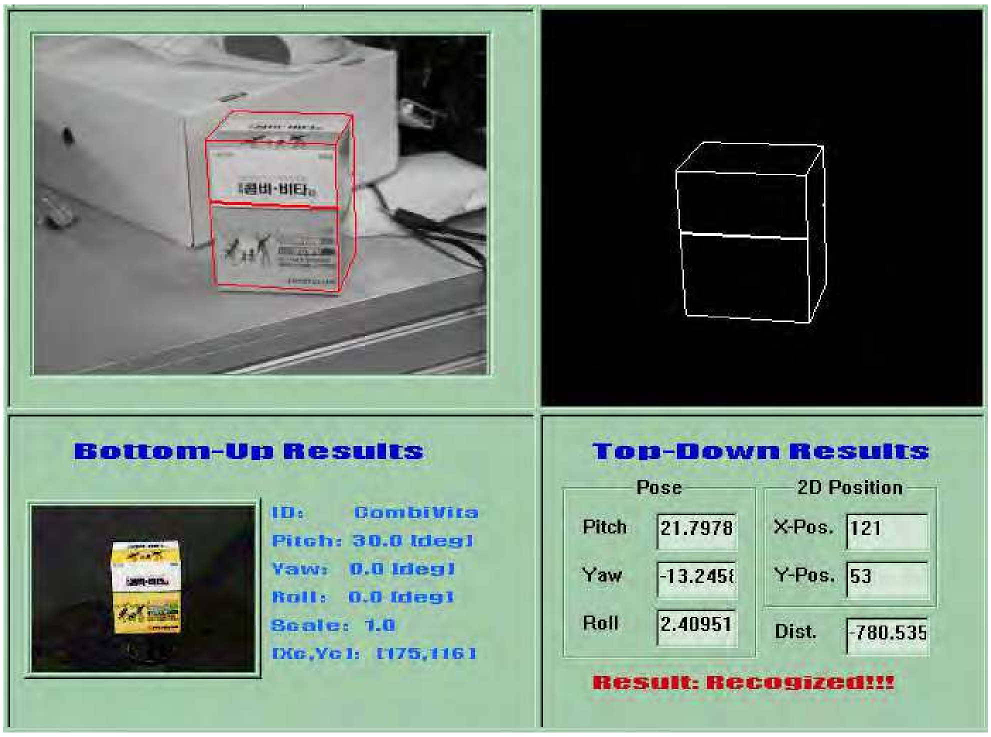
(a) (b) (c)

Figure 13. Parameter optimization by top-down process: (a) CAD model is overlaid with initial parameters, (b) after 10 iterations (c) after 40 iterations for the visible object

7. Experimental Results

In this paper, our main goal is to recognize man-made architectures such as building, bridge, container, and etc. using a FLIR camera. As an initial test, we experimented on a polyhedral object using a CCD camera. Then we evaluated the system on a FLIR dataset.

Figure 14 shows the overall interface of the target recognition system. This automatic target recognition system estimates the initial object parameters using scene and object DB. Then optimal parameter tuning is performed in top-down meaningful shape matching. From this result, the system makes a decision and feedbacks to the bottom-up process.



Bottom-Up Results

ID: CombiVita
Pitch: 30.0 [deg]
Yaw: 0.0 [deg]
Roll: 0.0 [deg]
Scale: 1.0
[Xc,Yc]: [175,116]

Top-Down Results

Pose
Pitch 21.7978
Yaw -13.245
Roll 2.40951

2D Position
X-Pos. 121
Y-Pos. 53
Dist. -780.535

Result: Recogized!!!

Figure 14. System interface-(upper left): input image with final result is overlaid, (upper right): rendered 3D CAD model generated from bottom-up and jump distribution, (lower left): bottom-up process result, (lower right): top-down process result which shows the optimal parameters

7.1 Test on visible database

First, we tested the algorithm for the objects captured using the CCD camera. We made a database for quantized views as explained. Figure 9 shows some results of the bottom-up process. We can get proper initial parameter values. Figure 13 shows the projection of a model with refined parameters by a top-down process for each object placed in the general environment. The overall computation time is 2 sec (0.5 sec for the bottom-up process) on the average under AMD 2400+.

7.2 Test on FLIR Database

The targets to recognize are shown in Figure 15. The sensor is FLIR Prism SP with resolution 320×240, NTSC interface. These models contain some background information which provide scene context. 3D CAD models are acquired by manual measurements.

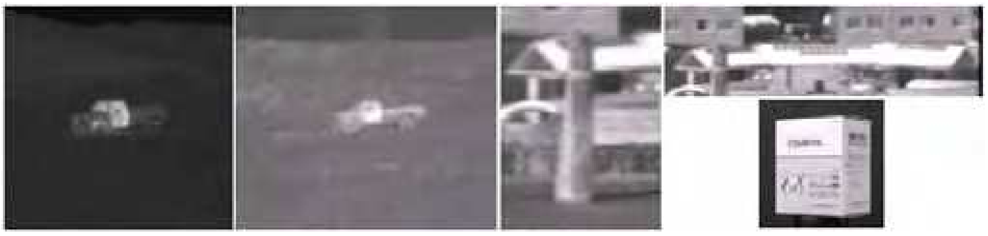


Figure 15. FLIR targets to recognize: cars, building, container, and tower

The test images are shown in Figure 16. They are composed of three types for the accurate performance evaluation for the practical use. The system has to recognize the targets in DB with high recognition rate and able to reject clutter objects or natural scenes.

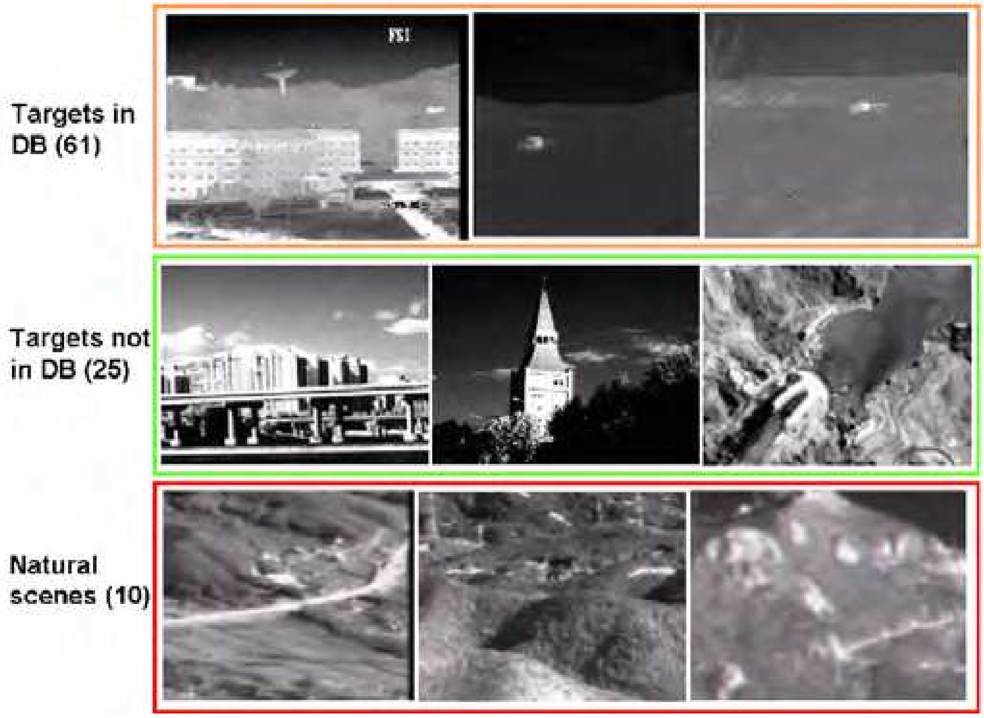


Figure 16. The composition of test images: targets in DB, targets not in DB, and natural scenes

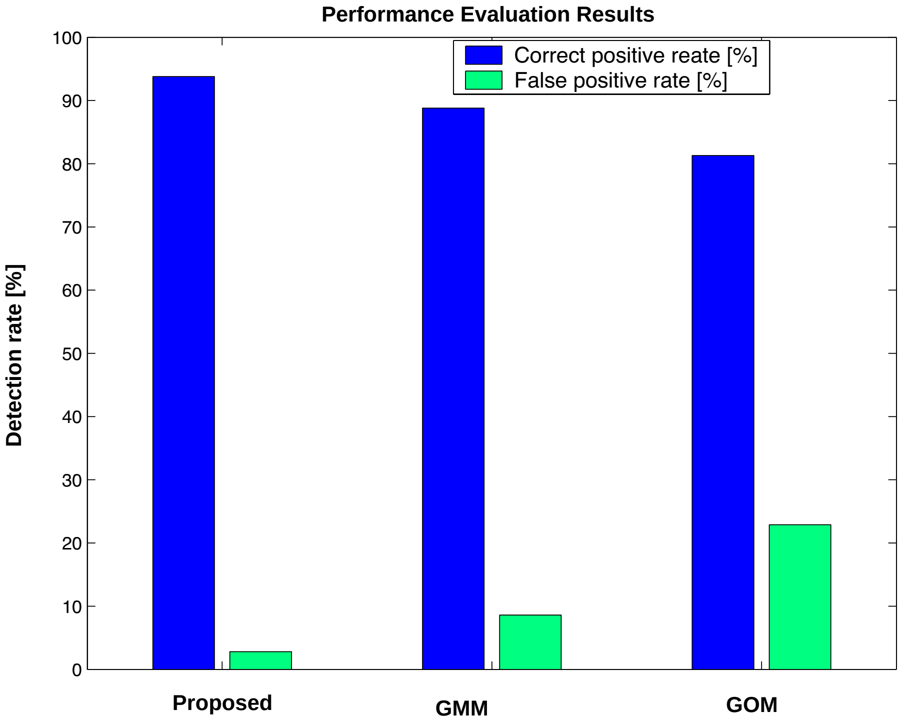


Figure 17. Evaluation of target recognition performance: the proposed method, GMM only, and GOM only

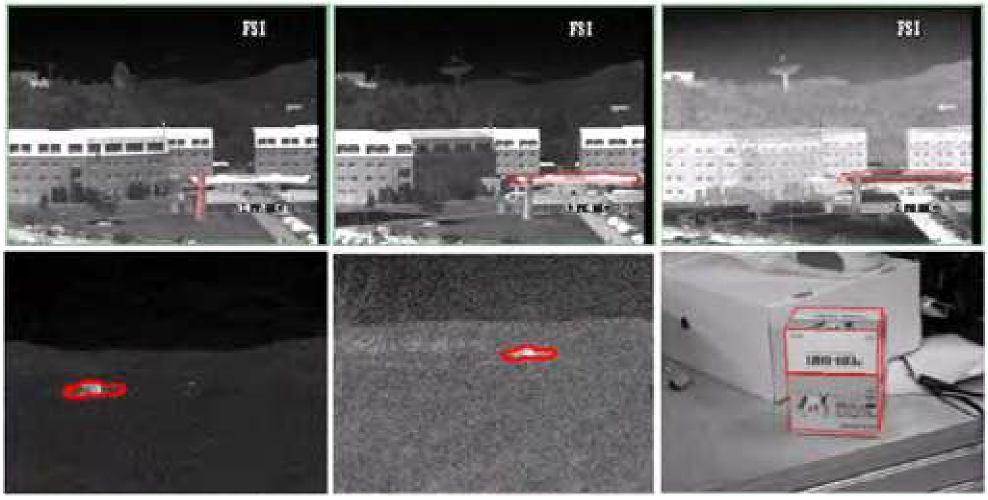


Figure 18. Successful recognition results

Figure 17 summarized our results compared with the methods of GMM only and GOM only. We used the performance measure as correct positive rate vs. false positive rate. In target recognition, the false positive rate is very important factor for practical system because false detections makes enormous damage. So, a good target recognition system has to high correct detection rate and very low false detection rate. During the performance comparison, we have the same bottom-up process with different top-down methods. We take all test images into consideration for the optimal parameter tuning. Our method outperforms the other two, with correct detection rate 93.75% and false detection rate only 2.85%. GOM-based method shows the worst performance. Figure 18 shows visual object recognition results for each object.

Figure 19 shows a typical failure case of the proposed system. The failures occurred from a bottom-up failure due to severe noise and a top-down failure due to low contrast.

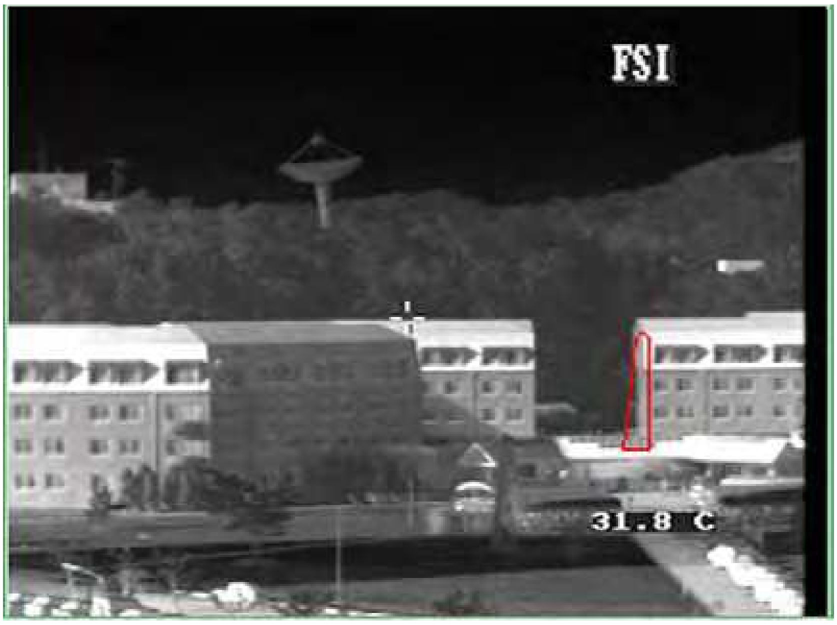


Figure 19. Failure case due to top-down fails due to low contrast

8. Conclusions

We propose a new object recognition paradigm based on the robust properties of the HVS, especially in scene context and 3D shape context information in a bottom-up and a top-down process. Furthermore, we also propose the cooperative feature map binding by utilizing both bottom-up and top-down processes and validate the system performance with various experiments. The test results on several images demonstrate efficiency in optimal matching as well as feasibility of the proposed recognition paradigm. The same paradigm will be extended to the general object recognition problem by changing the model representation.

9. Acknowledgments

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government(MOST)(No. M1-0302-00-0064), by the MIC for the project, “Development of Cooperative Network-based Humanoids’ Technology” of Korea.

10. References

Bar, M. (2004). Visual objects in context. *Nature Reviews: Neuroscience*, Vol. 5, 617-629.

Biederman, I. (1987). Recognition by Components: A Theory of Human Image Understanding. *Psychol. Review*, Vol. 94, No. 2, 115-147.

Borgelt, C. & Kruse, Z. (2001). *Graphical models: methods for data analysis and mining*. Wiley, New York, 1-12.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8, No.6, 679-698.

Desolneux, A.; Moisan, L. & Morel, J.M. (2004). Gestalt theory and computer vision. In *Carsetti A. Seeing, Thinking and Knowing*, Kluwer Academic Publishers, New York, 71-101.

Doucet, A.; Freitas, N.D. & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*, Springer, New York, 432-444, 3-13.

Edelman, S. & Bülthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, Vol. 32, 2385-2400.

Faugeras, O.D. & Hebert, M. (1986). The representation recognition, and locating of 3-D objects. *International Journal of Robotics Research,* Vol. 5, No. 3, 27–52.

Feldman, J. & Singh, M. (2005). Information along contours and object boundaries. *Psychological Review*, Vol. 112, No. 1, 243-252.

Fergus, R.; Perona, P. & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 264-271, Madison, Wisconsin, June.

Fiser, J.; Subramaniam, S. & Biederman, I. (2001). Size Tuning in the absence of spatial frequency tuning in object recognition. *Vision Research*, Vol. 41, No. 15, 1931-1950.

Green, P. (1996). *Reversible jump Markov Chain Monte Carlo computation and Bayesian Model Determination*, Champman and Hall, London.

Harris, C.J. & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of 4th Alvey Vision Conference*, 147-151, Manchester.

Kim, S. & Kweon, I.S. (2005). Automatic model-based 3D object recognition by combining feature matching with tracking. *Machine Vision and Applications*, Vol. 16, No. 5, 267-272.

Kumar, V.P. (2002). Towards trainable man-machine interfaces: combining top-down constraints with bottom-up learning in facial analysis. *Ph.D Thesis*, MIT.

Kuno, Y.; Ikeuchi, K. & Kanade, T. (1988). Model-based vision by cooperative processing of evidence and hypotheses using configuration spaces. *SPIE Digital and Optical Shape Representation and Pattern Recognition*, Vol. 938, 444-453.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, Vol. 30, No. 2, 77-116.

Lowe, D.G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, Vol. 31, No. 3, 355-395.

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Vol. 60, No. 2, 91-110.

Mikolajczyk, K. & Schmid, C. (2003). A performance evaluation of local descriptors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 774-781, Madison, Wisconsin.

Milanese, R.; Wechsler H. & Gil, S. (1994). Integration of bottom-up and top-down for visual attention using non-linear relaxation. *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 781-785, Seattle, USA , June.

Mundy, J. & Zisserman, A. (1992). *Geometric invariance in computer vision*, 335-460, MIT Press, Cambridge, MA.

Murase, H. & Nayar, S. (1995). Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, Vol. 14, 5-24.

Nichols, M.J. & Newsome, W. T. (1999). The neurobiology of cognition. *Nature*, Vol. 402, No. 2, C35-C38.

Parkhurst, D.; Law, K. & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, Vol. 42, 107-123.

Pasupathy, A. & Connor, C.E. (2001). Shape representation in area V4: position-specific tuning for boundary conformation. *Journal of Neurophysiology*, Vol. 86, No. 5, 2505-2519.

Peters, G. (2000). Theories of three-dimensional object perception - A Survey. In *Recent Research Developments in Pattern Recognition*, Transworld Research Network, Part-I, Vol. 1, 179-197.

Reisfeld, D.; Wolfson, H. & Yeshurun, Y. (1995). Context-free attentional Operators: the generalized symmetry transform. *International Journal of Computer Vision*, Vo. 14, No. 2, 119-130.

Ristic, B.; Arulampalam, S. & Gordon, N. (2004). *Beyond the Kalman filter: particle filters for tracking applications,* Artech House Publishers, London, 35-62.

Robert, C.P. & Casella, G. (1999). *Monte Carlo statistical methods*, Springer, New York.

Rothganger, F.; Lazebnik, S.; Schmid, C., & Ponce, J. (2004). Segmenting, modeling, and matching video clips containing multiple moving objects. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 914-921, Washington, DC, June.

Rothwell, C.A. (1993). Recognition using projective invariance, *Ph.D Thesis*, Oxford.

Schmid, C.; Mohr, R. & Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, Vol. 37, No. 2, 151-172.

Serre, T. & Riesenhuber, M. (2004). Realistic modeling of simple and complex cell tuning in the HMX model, and implications for invariant object recognition in cortex. *AIM*, MIT.

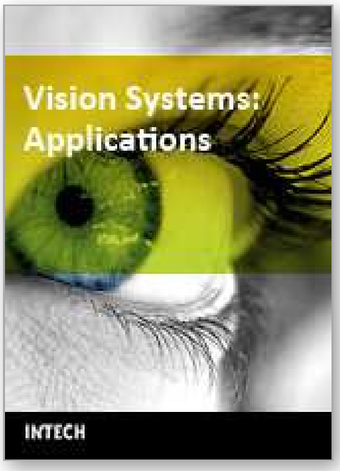
Siegel, M.; Kording, K.P. & Konig, P. (2000). Integrating top-down and bottom-up sensory processing by somato-dendritic interactions. *Journal of Computational Neuroscience*, Vol. 8, 161-173.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions: Biological Sciences* 29, Vol. 353, No. 1373. 1295-1306.

Tu, Z.; Chen, X.; Yuille, A. & Zhu, S.C. (2005). Image parsing: unifying segmentation, detection, and object recognition. (Marr Prize Issue, a short version appeared in ICCV 2003), *International Journal of Computer Vision*, Vol. 63, No. 2, 113-140.

VanRullen, R. (2003). Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology (Paris)* 97, 365-377.

Zhu, S.C.; Zhang, R. & Tu Z. (2000). Integrating bottom-up/top-down for object recognition by data driven Markov Chain Monte Carlo. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 738-745, Hilton Head, SC, June.



Vision Systems: Applications

Edited by Goro Obinata and Ashish Dutta

ISBN 978-3-902613-01-1

Hard cover, 608 pages

Publisher I-Tech Education and Publishing

Published online 01, June, 2007

Published in print edition June, 2007

Computer Vision is the most important key in developing autonomous navigation systems for interaction with the environment. It also leads us to marvel at the functioning of our own vision system. In this book we have collected the latest applications of vision research from around the world. It contains both the conventional research areas like mobile robot navigation and map building, and more recent applications such as, micro vision, etc.The fist seven chapters contain the newer applications of vision like micro vision, grasping using vision, behavior based perception, inspection of railways and humanitarian demining. The later chapters deal with applications of vision in mobile robot navigation, camera calibration, object detection in vision search, map building, etc.

INTECH

open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821