Pattern Recognition 43 (2010) 1116–1128

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Familiarity based unified visual attention model for fast and robust object recognition

Seungjin Lee*, Kwanho Kim, Joo-Young Kim, Minsu Kim, Hoi-Jun Yoo

Division of Electrical Engineering, School of Electrical Engineering & Computer Science, KAIST, 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 8 January 2009
Received in revised form 10 June 2009
Accepted 30 July 2009

Keywords:
Visual attention
Object recognition
Scene analysis

ABSTRACT

Even though visual attention models using bottom-up saliency can speed up object recognition by predicting object locations, in the presence of multiple salient objects, saliency alone cannot discern target objects from the clutter in a scene. Using a metric named familiarity, we propose a top-down method for guiding attention towards target objects, in addition to bottom-up saliency. To demonstrate the effectiveness of familiarity, the unified visual attention model (UVAM) which combines top-down familiarity and bottom-up saliency is applied to SIFT based object recognition. The UVAM is tested on 3600 artificially generated images containing COIL-100 objects with varying amounts of clutter, and on 126 images of real scenes. The recognition times are reduced by 2.7× and 2×, respectively, with no reduction in recognition accuracy, demonstrating the effectiveness and robustness of the familiarity based UVAM.

1. Introduction

Recently, local feature based object recognition approaches such as the SIFT [1,2] algorithm have grown popular due to their good invariance to size, rotation, and illumination when compared to traditional template based methods. However, the multiple transformations that are required by SIFT to achieve invariance require complex calculations. While runtimes vary depending on image content and size of the object database, SIFT currently cannot achieve real-time object recognition ($>$ 15 fps) on 640×480 pixel images on a modern PC. This limits its usefulness in real-time applications such as mobile robots.

Visual attention can be used to improve the runtime of object recognition by limiting analysis to regions likely to contain significant information. In fact, attention has been identified as a necessity for both human and machine vision. Due to the limited capacity of the brain, Neisser argued that a purely parallel model of vision is unfeasible [3]. Tsotsos also substantiates that claim by formally proving the NP-completeness of a parallel solution to the visual search task [4].

Bottom-up saliency based computational models of visual attention have been widely used to speed up object recognition. In [5], Itti et al. demonstrated a practical implementation of the bottom-up saliency map that was previously proposed by Koch and Ullman [6]. Several works have used this implementation of the saliency map to speed up object recognition tasks. Walther, Rutishauser et al. first demonstrated the usefulness of saliency-based attention for SIFT based object recognition, including the ability to perform unsupervised learning of objects from cluttered scenes [7,8]. In [9], Walther and Koch proposed a biologically plausible model of forming and attending proto-objects using bottom-up saliency. More recently, Hou and Zhang proposed a different method for saliency map generation that uses spectral residuals obtained by analyzing the log-spectrum of an input image [10]. This method was used by Meger et al. in [11] to construct a robot system employing attention based object recognition.

However, methods using only bottom-up saliency may not be optimal for tasks which can access a priori knowledge of the objects (i.e. robot navigation using pre-learned landmarks). Humans are known to speed up visual search by using prior knowledge of objects to attend to certain stimuli [12]. The cost of switching attention between such stimuli was measured by Walther and Li in [13]. At the cellular level, Fecteau and Munoz presented evidence that a combination of saliency and task relevance affect the firing of neurons [14].

Several top-down attention approaches that use pre-learned knowledge in object recognition were proposed. The approach by Itti's group used pre-learned characteristics of the target object to assign weights to the low level stimuli used to generate the saliency map [15,16]. Tsotsos et al. used feature direction, location, and abrupt onset and offset events as locational cues to bias selective tuning through the visual processing hierarchy [17]. Olivia et al. used statistical knowledge of the relationship between scene context and target objects to modulate attention [18]. In [19], Deco and Schürmann proposed a hypothesis–analysis loop in which the

* Corresponding author. Tel.: +82 42 350 5468; fax: +82 42 350 3410.
E-mail address: seungjin@eeinfo.kaist.ac.kr (S. Lee).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.07.014

spatial resolution of a region of interest (ROI) is progressively enhanced by top-down control.

In this paper, we propose "familiarity" as a metric for guiding top down attention. Familiarity is a measure of the resemblance of local features extracted from the input image to features of trained object models stored in a database. Features of high familiarity are seen as evidence of object existence, and are used to guide attention to locations likely to contain the corresponding object. An advantage of using familiarity over previous top down methods is that it does not require additional information other than the object database, which should be already available in an object recognition system.

Based on familiarity, the unified visual attention model (UVAM) that incorporates both bottom-up saliency and top-down familiarity is proposed. For bottom-up attention, Itti's saliency based visual attention model [5] is employed. The UVAM is applied to SIFT based object recognition to demonstrate its performance. Modifications are made to the conventional SIFT processing flow to facilitate information exchange between attention and recognition required to compute familiarity. The resulting reduction in recognition time greatly outweighs the overhead of the additional attention stages.

This paper is organized as follows. Section 2 will explain the UVAM, including details of the newly proposed familiarity based top down visual attention. In Section 3, the proposed model will be applied to a general purpose object recognition algorithm. Section 4 will summarize the performance of the UVAM including an analysis of the complementary nature of the bottom-up and top-down mechanisms. Finally, the conclusion will be given in Section 5.

2. Unified visual attention model

Unlike saliency, which is computed directly from the input image, familiarity is computed using the intermediate results of the object recognition process. Consequently, familiarity is only as effective as the quality of the object recognition results that are available. Initially, the UVAM includes a preliminary object recognition phase that performs quick feature extraction on the input image. This essentially provides a low resolution snapshot of the input feature space similar to the hierarchical feature extraction approach of [19]. During the detailed object recognition phase, the familiarities of newly extracted features are continuously reflected in the top-down attention, thus forming an attention–recognition feedback loop that continuously improves both attention and recognition accuracy.

The outline of the proposed UVAM is shown in Fig. 1. The top-down and bottom-up components of the UVAM can be divided into two stages: the feed-forward attention stage of the left hand side, and the attention feedback loop of the right hand side. The feed-forward attention stage provides a preliminary estimation of the location of trained objects before starting the detailed object recognition. The attention feedback loop updates this estimation later based on the results of detailed object recognition on each selected ROI.

The bottom-up saliency map ($\mathcal{S}$-map) [5] is calculated once during the feed-forward attention stage. In contrast, top-down familiarity is calculated once during the feed-forward attention stage to obtain the feed-forward familiarity map (FF $\mathcal{F}$-map), and then repeatedly during the attention–recognition feedback loop to obtain the feedback familiarity map (FB $\mathcal{F}$-map). The $\mathcal{S}$-map and the two $\mathcal{F}$-maps are combined into the unified attention map ($\mathcal{UA}$-map), which is used to select the ROI for detailed object recognition.

2.1. Saliency based bottom-up attention

The saliency based visual attention model [5] is a biologically inspired visual attention algorithm for identifying conspicuous locations in a scene. The model is based on the previous work of Koch and Ullman [6], which modeled selective attention in primate vision as a competition between salient low-level features in the visual stimuli. The model uses the low-level features, color, intensity, and orientation, to generate a saliency map ($\mathcal{S}$-map) which represents the saliency of each location in the input image by a scalar quantity.

However, using the bottom-up $\mathcal{S}$-map alone for guiding attention may result in sub-optimal results depending on the clutter content of the scene. The $\mathcal{S}$-map is most accurate for scenes in which the object of interest is conditioned for visual pop-out [20]. Visual pop-out occurs when the target object can be distinguished from distractors by a single feature type, in which case a dominant peak at the location corresponding to the object is observed on the $\mathcal{S}$-map. Performance is degraded, however, when the scene includes distractors which have higher saliency than the target object.

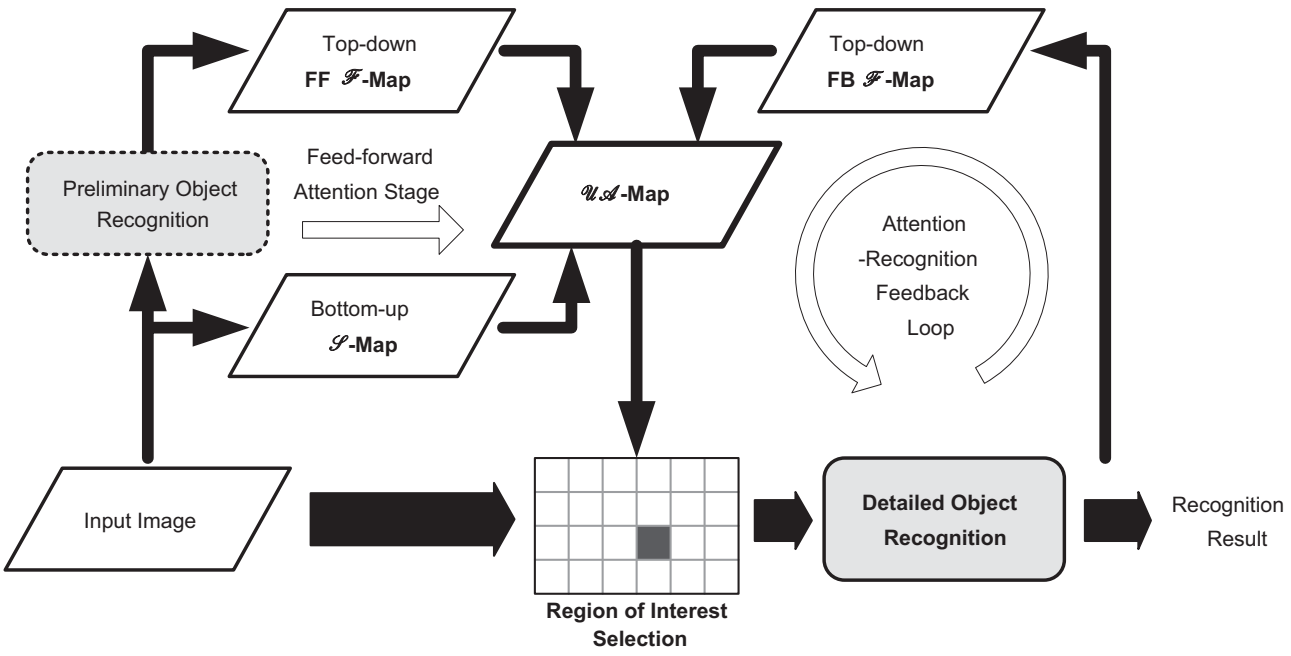

Fig. 1. Outline of the unified visual attention model.

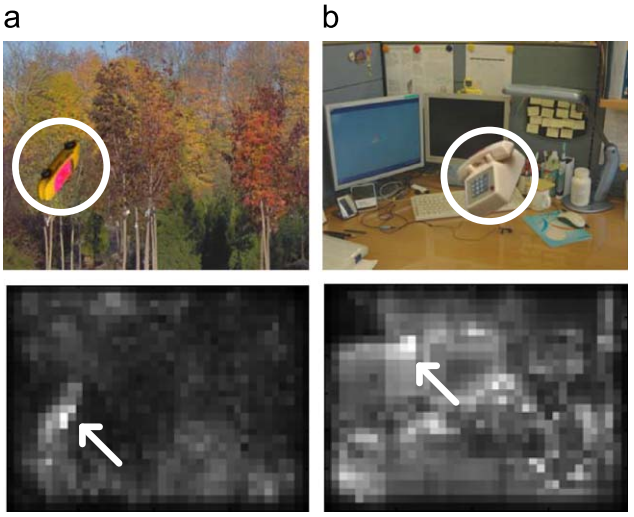

Fig. 2. Usefulness of bottom-up saliency-based attention for (a) a scene conditioned for visual pop-out and (b) a scene with salient background clutter. The circles mark target objects in the scene, and the arrows mark the point of highest saliency in the scene.

The two situations are illustrated in Fig. 2. In Fig. 2(a), the bright yellow and pink segments that compose the target object, a toy car, make the target object stand out from the non-salient forest background as is clearly indicated by the bright blob in the $\mathcal{S}$-map. However, in Fig. 2(b), the target object, a beige colored telephone, is not the most salient object in the scene due to the cluttered office background. As a result, the target object has lower attention priority than a large portion of the background.

2.2. Familiarity based top-down attention

Psychological experiments have shown that human vision exhibits an attentional bias towards familiar objects. For example, in [21], subjects asked to identify motion of familiar and unfamiliar two-letter strings displayed preferential processing of familiar items. The study found that this preferential processing occurs as a result of a sub-conscious process rather than through the conscious intent of the subject. Another experiment showed that visual search could be speeded up by pre-cueing the target location with a shape held in memory [22]. These results show that attention is directed towards familiar objects, even when there is no explicit intention of finding those objects.

Familiarity is calculated using intermediate results of the feature based object recognition process. In this study, matching results of individual SIFT keypoints and clusters of SIFT keypoints are used for the calculation of familiarity. When an individual query keypoint from the input image (k_q) is matched to a keypoint in the object database (k_m), the distance measure between the two keypoints (d), is returned as the result. Similarly, when two or more keypoints k_{q1},k_{q2},…,k_{qn} are clustered, the distance measure Δ_{ij} is calculated between each possible combination of pairs of keypoints. In both cases, the smaller the distance measure, the higher the probability of a true positive match. Hence, the familiarity of individual keypoints and clusters of keypoints are defined to be inversely proportional to their respective distance measures. The following two subsections describe the definition of familiarity of individual keypoints, $\mathcal{F}_{\text{keypoint}}$, and familiarity of keypoint clusters, $\mathcal{F}_{\text{cluster}}$.

2.2.1. Familiarity of keypoints

The familiarity of an individual keypoint should represent the similarity between that keypoint and keypoints in the object database. It is calculated using the Gaussian function as

$$\mathcal{F}_{\text{keypoint}} = \exp(-d^2/(2\sigma_k^2)). \tag{1}$$

This basically assigns an inversely proportional relationship between familiarity and the distance measure $d = |\mathbf{d}_q - \mathbf{d}_m|$, which is the Euclidean distance between the 128 dimensional SIFT descriptor vectors [2] of k_q, the query keypoint, and k_m, its closest matching keypoint in the object database. Since SIFT keypoint descriptor vectors are normalized to 1 and have positive valued elements, d lies in the range between 0 (exact match) and $\sqrt{2}$ (orthogonal). The Gaussian function assigns high familiarity to keypoints with small d, and low familiarity to keypoints with large d. The constant σ_k determines the selectivity of the Gaussian function and thus the range of distances that are assumed to be familiar.

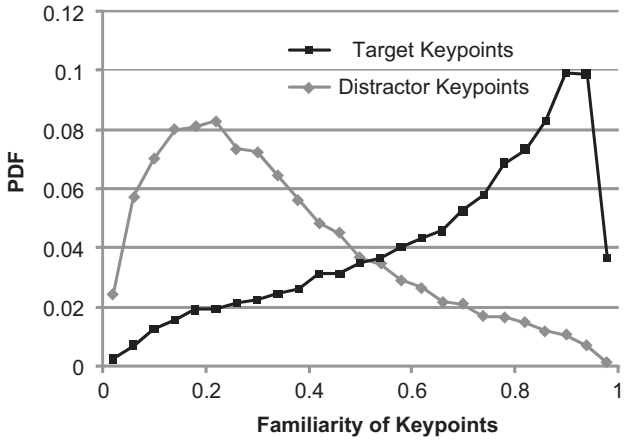

Fig. 3. PDF of the familiarity of keypoints extracted from target objects, and that of keypoints extracted from distractor objects.

The value of σ_k must be selected to maximize the selectivity between "target keypoints" and "distractor keypoints". In the images tested for this study, on average only 5% of the total extracted SIFT keypoints comes from target objects, while the remaining 95% is from distractors. This implies that $\mathcal{F}_{\text{keypoint}}$ must have sufficiently high discriminability between the "target keypoints" and the "distractor keypoints" in order to prevent the familiarity of the target keypoints from being obscured by that of the distractor keypoints. Based on the PDF of the distance measure d of the target keypoints and the distractor keypoints, σ_k was chosen to be 0.25 to maximize the ratio between the expected value of familiarity for target keypoints and distractor keypoints, or $E(\mathcal{F}_{\text{target-keypoint}})/E(\mathcal{F}_{\text{distractor-keypoint}})$. Fig. 3 shows the resulting PDFs of $\mathcal{F}_{\text{keypoint}}$ for target keypoints and distractor keypoints.

2.2.2. Familiarity of keypoint clusters

The familiarity of clusters of keypoints, used for FB $\mathcal{F}$-map generation, is defined as follows:

$$\mathcal{F}_{\text{cluster}} = \begin{cases} \exp(-\Delta_{ij}/(2\sigma_c^2)), & \text{cluster size} = 2 \\ -2, & \text{cluster size} > 2 \end{cases}. \tag{2}$$

Δ_{ij} is the distance measure used for keypoint clustering (see Section 3) which measures the likeliness that two keypoints, i and j, are part of the same target object. Ideally, Δ_{ij} is equal to 0 for keypoints originating from the same object but is larger for keypoints originating from random clutter. The Gaussian function assigns high familiarity to keypoint clusters with small Δ_{ij}.

$\mathcal{F}_{\text{cluster}}$ assumes a positive value only when the cluster size is 2. For clusters with more than two keypoints, the value of $\mathcal{F}_{\text{cluster}}$ is −2. From our test images it is found that clusters of three or more

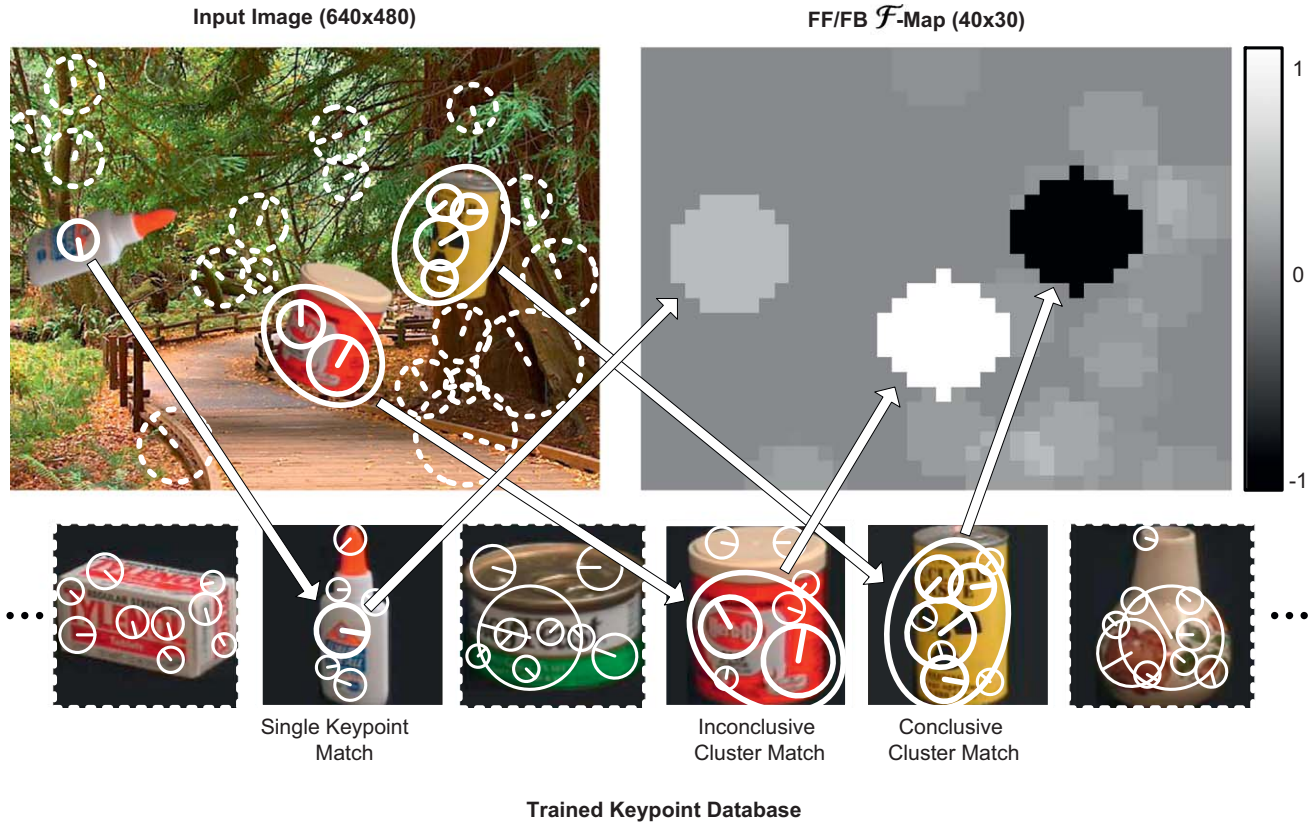

Fig. 4. Conceptualization of the $\mathcal{F}$-map generation process. Single keypoints matches and inconclusive cluster matches (two keypoints) are viewed as evidence of a target object and are represented as positive valued ellipses on the familiarity map. Conclusive cluster matches (three or more keypoints) are represented as negative valued ellipses on the $\mathcal{F}$-map to inhibit further analysis.

keypoints have negligible false positive rates, and thus do not require further analysis. Therefore, when clusters of three or more keypoints are found, redundant calculations are avoided by preventing further detailed analysis of the object. The $\mathcal{F}_{\text{cluster}}$ value of -2 achieves this by canceling any positive values of the $\mathcal{S}$-map and $\mathcal{F}_{\text{keypoint}}$.

2.2.3. Familiarity map ($\mathcal{F}$-map) generation

The FF $\mathcal{F}$-map and FB $\mathcal{F}$-map are generated using $\mathcal{F}_{\text{keypoint}}$ and $\mathcal{F}_{\text{cluster}}$, respectively. Since $\mathcal{F}_{\text{keypoint}}$ and $\mathcal{F}_{\text{cluster}}$ are scalar values, a method is needed for projecting them onto the 2D $\mathcal{F}$-maps, whose values correspond to the familiarity of rectangular (16×16 in our case) pixel regions of the input image. Optimally, the projected shape should match the shape of the actual object. In systems employing bottom-up saliency, several methods have been proposed to predict object shape using only bottom-up information. These include simple thresholding of the $\mathcal{S}$-map [10], finding homogeneous regions in the feature map that contributed most to the attended location [7,8], and grouping using motion [23]. However, a more accurate representation of object shape is possible if prior knowledge about the target object is used. In our approach, the object model stored in the object database is used to approximate the object shape. The object shape is approximated using the inscribed ellipse of the bounding box of the object model. This is simpler and more computationally efficient than using the exact object outline, while being sufficiently accurate for our needs.

Since the pose of the object in the image is different from that of the object database, we must first calculate the pose of the target object relative to that of the object database. Using the pre-trained information in the object database, the pose $p=\{x, y, S, \theta\}$ of the predicted object is first calculated from the keypoint information assuming a similarity transform, where x and y are the coordinates of the object center, S is the size of the object, and θ is the orientation of the object. After the pose is estimated, the familiarity value is added to pixels of the $\mathcal{F}$-map that lie within the inscribed ellipse of the bounding box of the predicted object, as shown in Fig. 4.

The FF and FB $\mathcal{F}$-maps are defined as

$$\text{FF } \mathcal{F}\text{-map}(x,y) = \sum_{i \in \text{keypoints}} ellipse_i(x,y)\mathcal{F}_{\text{keypoint}_i} \tag{3}$$

and

$$\text{FB } \mathcal{F}\text{-map}(x,y) = \sum_{i \in \text{clusters}} ellipse_i(x,y)\mathcal{F}_{\text{cluster}_i}, \tag{4}$$

respectively, where $ellipse_i(x,y)$ is an indicator function defined as follows:

$$ellipse_i(x,y) = \begin{cases} 1, & (x,y) \text{ lies inside ellipse defined} \\ & \text{by pose of keypoint or cluster } i \\ 0, & (x,y) \text{ lies outside ellipse defined} \\ & \text{by pose of keypoint or cluster } i \end{cases}. \tag{5}$$

The FF $\mathcal{F}$-map generation requires a dedicated preliminary object recognition on the input image, which causes execution time overhead. In order to minimize the additional processing time for the preliminary object recognition stage, the spatial resolution of the input image is reduced by a reduction factor λ, and a matching error ε is introduced in its keypoint matching step as shown in Fig. 5. The

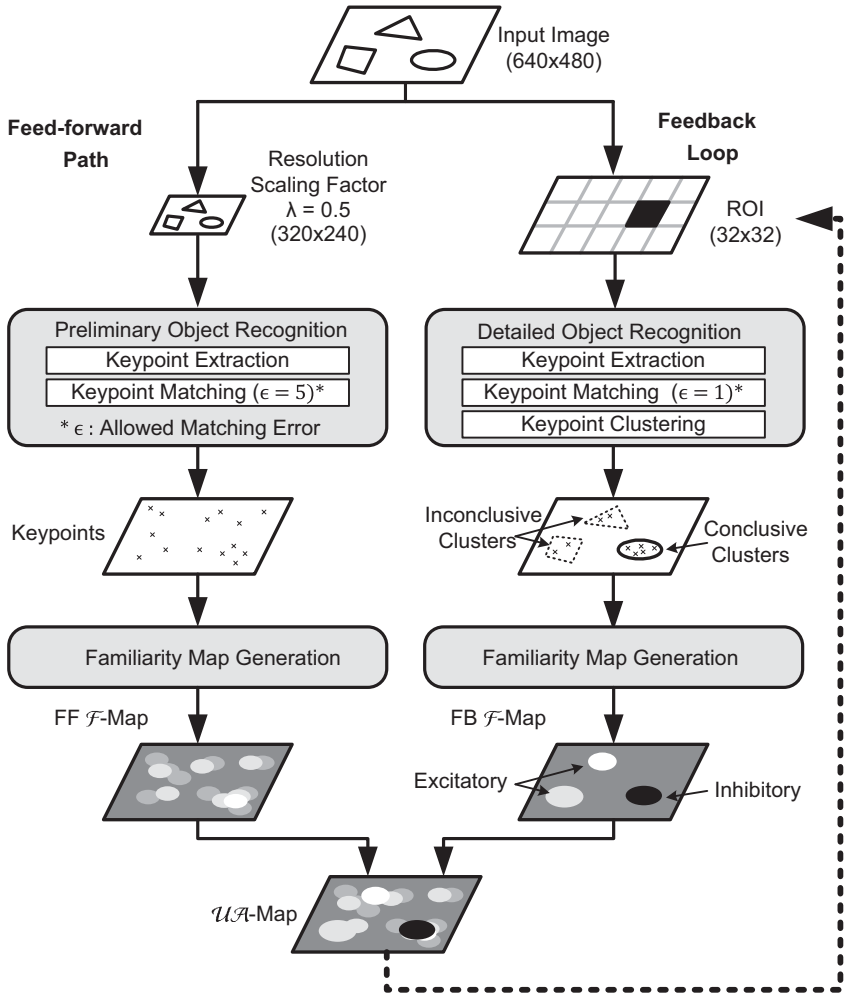

Fig. 5. Overview of FF and FB $\mathscr{F}$-map generation. The feed-forward process executes a reduced version of object recognition on the entire reduced resolution input image. The feedback loop executes detailed object recognition on a small ROI of the full resolution input image.

introduction of these parameters trades off prediction accuracy with computation speed. Excessively small λ may make keypoints that originate from small details in the input image undetectable due to the reduced resolution. Meanwhile, increasing ε may result in some of the target keypoints being misclassified as clutter during the keypoint matching process. Using $\lambda = 0.5$ resulted in a 60% reduction in average number of keypoints and using $\varepsilon = 5$ resulted in 20% of target keypoints being misclassified. With these values of λ and ε, the execution time for the preliminary object recognition was reduced to less than 1/10 of detailed object recognition (Fig. 6).

The FB $\mathscr{F}$-map is generated using the keypoint clustering result of each iteration of detailed object recognition. In contrast to the preliminary object recognition, detailed object recognition executes with higher resolution ($\lambda = 1$) and matching accuracy ($\varepsilon = 1$). Due to the large number of keypoints (and thus background clutter) that are detected during detailed object recognition, only the clusters of keypoints are considered for familiarity feedback. The purpose of the familiarity feedback mechanism is twofold as shown in Fig. 6. One is to identify and assist the selection of familiar regions in the image. The other is to inhibit the selection of regions that have already been concluded to contain a trained object. This inhibition process allows the detailed object recognition process to move on to "fresh" regions once an object is positively identified in order to reduce the total execution time.

2.3. ROI selection

The most common method of attending to locations in bottom-up attention approaches is to apply winner take all (WTA) on the $\mathscr{S}$-map, then use some kind of inhibition of return mechanism [5,7,8,10,23]. The unit of attention in those cases can be simple discs [5], or the shape of the estimated object outline [7,8,10,23]. We take a similar approach except we use the $\mathscr{UA}$-map, which is the sum of the $\mathscr{S}$-map and FF and FB $\mathscr{F}$-maps. Additionally, we use predefined tile shaped ROIs as the unit of attention. The predefined ROIs are usually much smaller than the actual object outline which allows objects to be recognized without analyzing the entire object region. In conjunction with the inhibitive familiarity feedback which was explained previously, this enables some reduction in execution time.

Our model divides a 640×480 pixel input image into 300 (15 rows by 20 columns) 32×32 pixel ROIs for the detailed recognition stage. The result of bottom-up attention, the $\mathscr{S}$-map, and the results of top-down attention, the FF and FB $\mathscr{F}$-maps, are added together to obtain the $\mathscr{UA}$-map. For each iteration of the attention–recognition feedback loop, the ROI that corresponds to the point of maximum value in the $\mathscr{UA}$-map is selected for detailed analysis. ROIs that were previously selected are excluded from subsequent iterations of the loop.

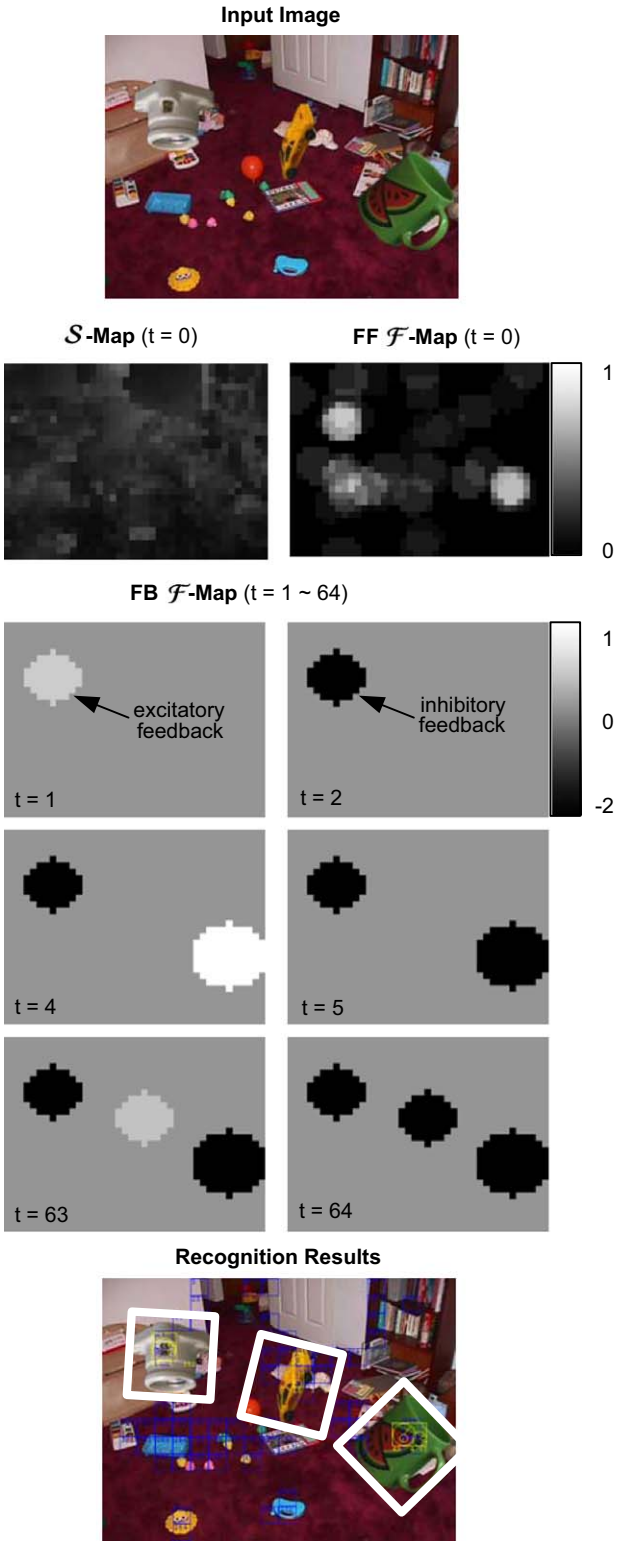

Fig. 6. Example of the operation of the UVAM on a scene with many salient distracters. Three target objects are successfully recognized after 64 iterations of the attention–recognition feedback loop. The FB $\mathcal{F}$-map is shown for iterations 1–64.

3. Fast and robust object recognition with unified visual attention

Previous attention based systems used SIFT [7,8,15,16,18] as well as biologically inspired methods that explicitly attempted to model the cortical structure of the human brain [9]. Recently, a biologically inspired object recognition system by Serre et al. was shown to outperform SIFT in classification of generalized categories [24]. However, SIFT is widely used in recognizing specific instances of objects, which is required for many tasks such as landmark recognition. In addition, the distinctiveness of SIFT features makes them suitable for calculating familiarity as shown in Fig. 3.

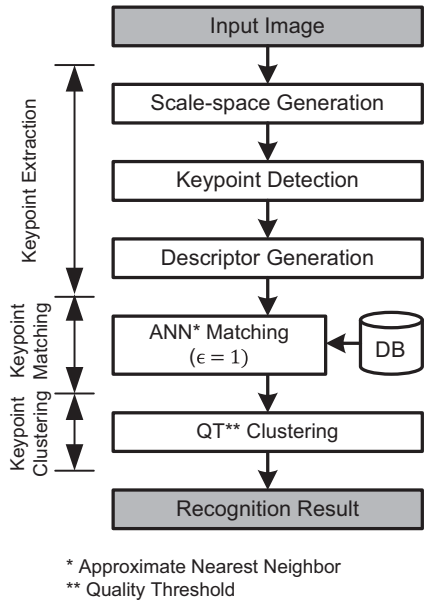

Fig. 7. SIFT based object recognition without visual attention.

In this work, we use SIFT as the base recognition system. SIFT can be divided into three main steps as shown in Fig. 7: keypoint extraction, keypoint matching, and keypoint clustering. The keypoint extraction stage extracts SIFT keypoints which encode the location, size, and texture information of features in the input image. During the keypoint matching stage, these keypoints are matched to their approximate nearest neighbor (ANN) in the database which stores the keypoints of trained objects. Keypoints that are likely to have originated from the same trained object, as explained in Section 3.5, are then clustered together in the keypoint clustering stage.

In this section, we analyze the execution times of each stage of the SIFT object recognition algorithm. After that, the proposed unified visual attention model is integrated into the reference SIFT based object recognition system. The modifications made to each step and their effects on performance are examined.

3.1. Object recognition without visual attention

The execution times of each stage of the reference SIFT object recognition before visual attention is integrated are measured and the contributions of each stage to the total execution time are evaluated. Fig. 8 shows the average execution times of each stage in SIFT based object recognition for 3600 synthesized images classified into three groups according to the number of keypoints; high (>1400), medium (900–1400), and low (<900). According to the analysis, the descriptor matching step is the most time consuming step primarily due to the large ($>40{,}000$ keypoint) database used. The keypoint extraction stage, which is composed of scale-space generation, keypoint detection, and descriptor generation, takes relatively short time. The time required for keypoint clustering is less than 1% of the total execution time and is not shown on the graph.

In addition to the relative contributions of each stage of object recognition to execution time, Fig. 8 shows that execution time

is highly dependent on the number of keypoints in the image. Specifically, descriptor generation and keypoint matching take time approximately proportional to the number of keypoints that are analyzed while the scale-space generation and keypoint detection stages take constant time regardless of the image contents. Based on these observations, the execution time of object recognition t_0 can be approximated by the linear equation:

$$t_0 = \alpha + \beta N, \tag{6}$$

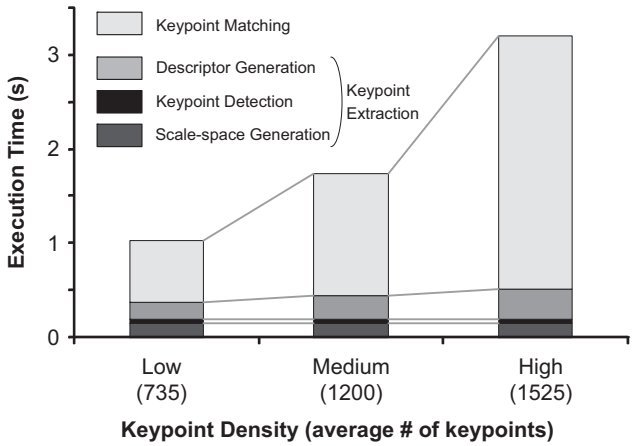

Fig. 8. Average execution times of each stage of SIFT feature extraction and matching without visual attention for scenes with low, medium, and high keypoint density. The contribution of keypoint clustering to total execution time is negligible and is not shown.

where α is the constant execution time independent of image contents, β is the multiplication coefficient, and N is the number of keypoints. Here, the keypoint dependent term βN accounts for 82–94% of the total execution time. Therefore, descriptor generation and especially keypoint matching should be restricted to as few keypoints as possible in order to minimize the execution time.

3.2. Applying the unified visual attention model to object recognition

The unified visual attention model (UVAM) needs to meet two requirements for its effectiveness in object recognition. One is the minimization of the number of keypoints subject to descriptor generation and keypoint matching. The other is the minimization of overhead imposed by the additional visual attention processes. Eq. (6) can be generalized to include the effects of visual attention as

$$t_A = \alpha + \gamma\beta N + \tau, \tag{7}$$

where t_A is the execution time of object recognition with visual attention, and γ and τ denote the keypoint reduction factor and attention overhead, respectively. In order to achieve fast object recognition, the UVAM should minimize the keypoint reduction factor γ without introducing significant attention overhead τ.

Fig. 9 shows the object recognition flow with the UVAM. The preliminary object recognition, shown on the left-hand side of Fig. 9, shares intermediate results with the detailed object recognition process of the right hand side in order to minimize the visual attention overhead τ. The preliminary object recognition required for FF $\mathcal{F}$-map generation uses the results of scale-space generation and keypoint detection on the original image shown at the top of Fig. 9, instead of operating on a separate image scaled down by

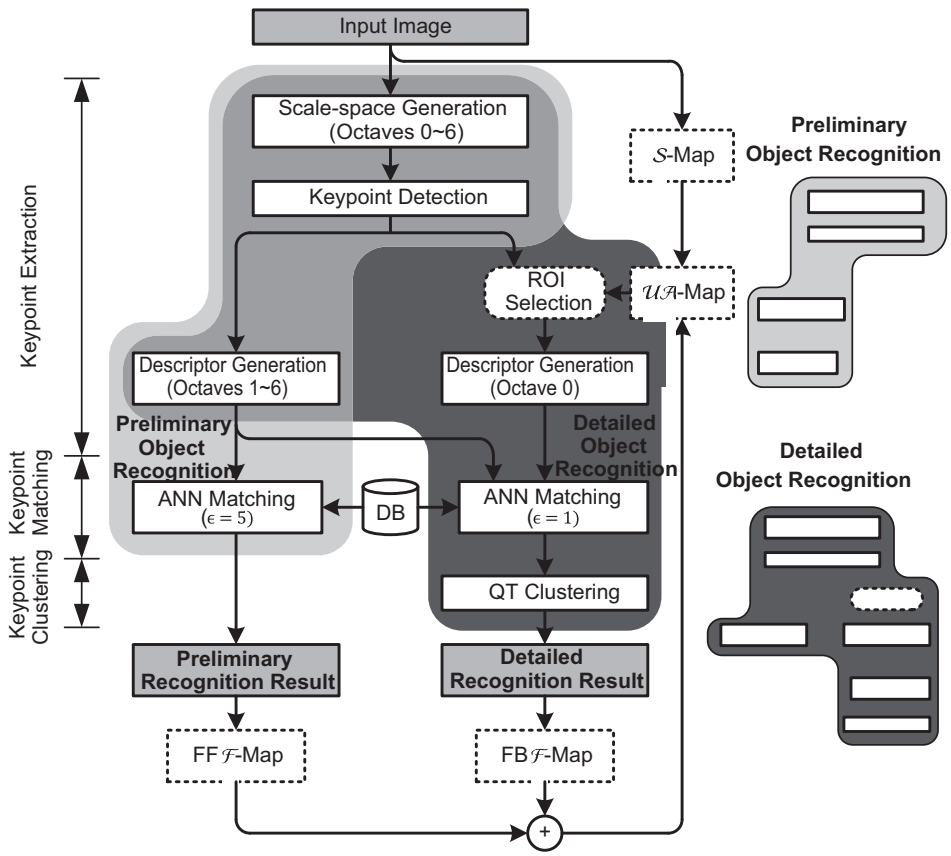

Fig. 9. SIFT based object recognition flow with the UVAM applied. The light gray region depicts the reduced flow for feed-forward familiarity map generation, the dark gray region the flow for detailed object recognition, and the medium gray region the steps that are shared between the two flows.

$\lambda = 0.5$ as described in Fig. 5. The equivalent effect of setting $\lambda = 0.5$ can be achieved by considering keypoints of scale ≥ 2 (or octaves ≥ 1). The descriptors for keypoints of scale ≥ 2 generated during this stage can be saved and reused for detailed object recognition, further reducing the overhead of visual attention.

During the detailed object recognition stage, detailed matching ($\varepsilon = 1$) is confined to keypoints that are located within ROIs selected by the UVAM, as shown in the right-hand side of Fig. 9. Here, SIFT descriptors are generated only for keypoints belonging to octave 0, since descriptors for keypoints of octaves 1–6 are previously calculated during the preliminary object recognition stage. The number of selected ROIs should be reduced to minimize the keypoint reduction factor γ.

In the following subsections each step of object recognition will be explained in detail with modifications introduced by the UVAM.

3.3. Keypoint extraction

The scale-space generation and keypoint detection steps of SIFT are executed once for both preliminary and detailed object recognition. Since preliminary object recognition analyzes the entire image, descriptors are calculated for all keypoints with scale ≥ 2. Descriptors for keypoints of scale < 2 are calculated only if the keypoint is located within the selected ROIs.

3.4. Keypoint matching

It is crucial to minimize the execution time of keypoint matching since it takes the longest time to execute among the object recognition steps as shown in Fig. 8. For SIFT keypoints, nearest neighbor matching using sophisticated search structures such as kd-trees exhibit poor performance [25] due to the high dimensionality (128) of the descriptor vectors. Fortunately, approximate methods such as the randomized neighborhood graph (RNG*) [26] can be used to achieve much higher speeds at the cost of introducing a small error into the search process.

In the RNG* method, the parameter ε, which is the same as the matching error previously mentioned in Section 2.2, is used to control the tradeoff between speed and accuracy. For a positive value of ε, the RNG* method guarantees that the Euclidean distance between the query vector and the returned approximate nearest neighbor vector, which may or may not be the true nearest neighbor, is smaller than $(1+\varepsilon)$ times the distance between the query vector and the true nearest neighbor. Fig. 10 shows keypoint matching accuracy and execution time as a function of ε for the $>40{,}000$ keypoint COIL-100 [27] database used in our experiments. With increase of ε, matching accuracy decreases linearly, but execution time decreases exponentially. The decrease in accuracy is especially small for target keypoints, which are of interest in this study.

Two values of ε are used for keypoint matching depending on whether the emphasis is on accuracy or on speed. Based on observations of Fig. 10, $\varepsilon = 1$ is used for the detailed object recognition, and $\varepsilon = 5$ is used for the preliminary object recognition. Choosing $\varepsilon = 1$ provides 99.9% matching accuracy for target keypoints with just 23% of the execution time of exact nearest neighbor search. Choosing $\varepsilon = 5$ results in matching accuracy of only 80% but reduces execution time to less than 1% of the exact case, and is thus suitable for preliminary object recognition which requires a quick keypoint matching.

3.5. Keypoint clustering

The goal of keypoint clustering is to obtain a cluster of keypoints that together predict the existence of an object and its pose in the image. The pose p is represented as $p = \{x, y, S, \theta\}$ as explained in Section 2.2.3. To obtain keypoint clusters, Lowe [2] uses a voting scheme based on the Hough transform [28] together with an affine transformation model using the least-squares method [29]. Although the Hough transform is computationally efficient, its binning based clustering method is not suitable for calculating familiarity which requires a method of evaluating the level of familiarity for inconclusive object matches.

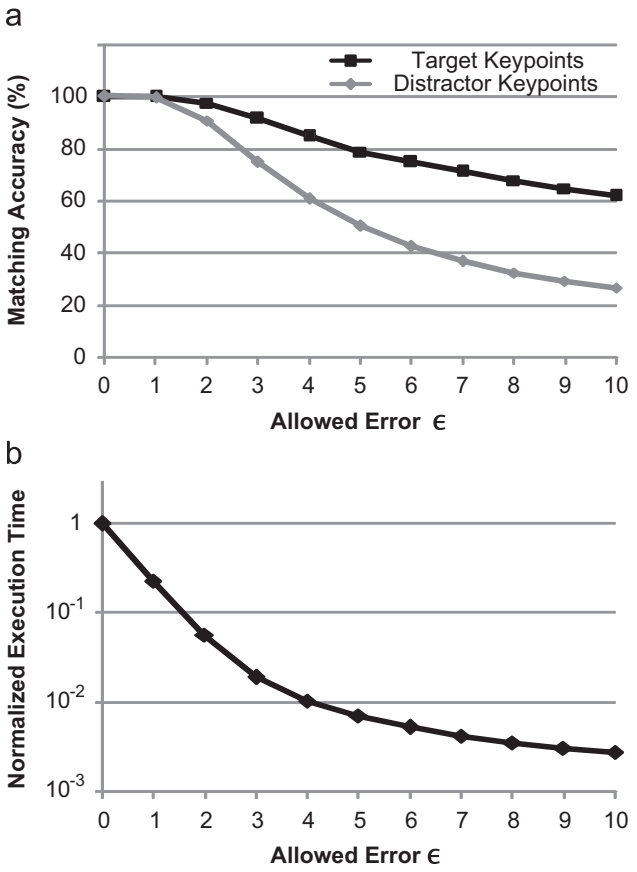

Fig. 10. (a) Percentage of correct matches and (b) execution time of keypoint matching using approximate nearest neighbor search with varying values of ε when compared to exact nearest neighbor search ($\varepsilon = 0$). Target keypoints are more tolerant to higher values of ε than distractor keypoints.

Quality threshold (QT) clustering [30], which is simple yet effective for obtaining clusters of high quality, is applied. The quality of a cluster C is quantified by its diameter D, defined as $D = \max_{i,j \in C}\{\Delta_{ij}\}$, where i and j are keypoints in cluster C, and Δ_{ij} is the distance measure between two keypoints. QT clustering ensures the quality of its clusters by limiting their diameters below a threshold D_{th}.

In this study, the distance measure Δ_{ij} between keypoints i and j is defined using the errors between the object poses $p_i = \{x_i, y_i, S_i, \theta_i\}$ and $p_j = \{x_j, y_j, S_j, \theta_j\}$ predicted by the keypoints:

$$\Delta_{ij} = \frac{\delta_{xy}}{S_{avg}} + \frac{\delta_s}{S_{avg}} + \frac{\delta_\theta}{\pi},$$

$$\delta_{xy} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

$$\delta_s = |S_i - S_j|, \quad \delta_\theta = |\theta_i - \theta_j|, \quad S_{avg} = (S_i + S_j)/2, \tag{8}$$

where δ_{xy} is the error between object locations, δ_s is the error between sizes, δ_θ is the error between orientations, and S_{avg} is the average of the predicted object size. The normalization of the error terms is necessary since they have different units and thus different ranges. For keypoints that originate from the same object, Δ_{ij} should

ideally equal zero. However, due to image noise and possible 3D rotations that cannot be predicted by the keypoints, we must allow for some errors between the predictions of the keypoints. By setting $D_{th} = 0.75$, an average of 25% error is allowed for each pose parameter. Increasing this threshold will result in higher true positive rates at the cost of higher false negative rates.

QT clustering is applied to objects that have been implicated by at least two keypoints. Each resulting cluster consisting of at least three keypoints is classified as conclusive object matches. This lower bound, which was also used in [2] for Hough transform clusters, provides accurate matches with a low rate of false matches even in the presence of background clutter. For each positive object match, the pose of the object is estimated as the average of the poses estimated using each individual keypoint.

4. Performance evaluation

A quantitative evaluation is carried out on 3600 test images generated using objects from the COIL-100 library. In addition, tests are carried out on two separate sets of natural images to further verify the robustness of the system. For each of the three test image sets, the object recognition system is trained using target object images taken at 30° viewpoint increments. Test image resolution is 640×480 pixels for the generated images, and 1280×960 for the natural images.

4.1. Performance of visual attention model

In order to accurately measure its object recognition performance, a large set of images containing trained objects with controlled amounts of background clutter is required. 3600 images are created by combining objects from the COIL-100 library with 12 natural background images containing varying amounts of detailed textures and salient objects as shown in Fig. 11. For each background image, 300 images are synthesized with one to three trained objects of randomized locations, sizes, and orientations. The keypoint database is constructed using images of target objects taken at 30° increments. To prevent template matching, only images from views that are not employed in constructing the keypoint database are used to generate the test images.

We measure execution time as the time taken to detect and localize all target objects in an input image. Keypoint count is the number of SIFT keypoints that are analyzed in detail (with matching error $\varepsilon = 1$). Fig. 12 compares the average keypoint count and execution times for object recognition using the following four configurations of visual attention.

1. No attention: all ROIs are analyzed.
2. Bottom-up saliency: ROI selection prioritized by $\mathscr{S}$-map.
3. Top-down familiarity: ROI selection prioritized by FF and FB $\mathscr{F}$-maps.
4. UVAM: ROI selection prioritized by $\mathscr{UA}$-map ($\mathscr{S}$-map, FF $\mathscr{F}$-map and FB $\mathscr{F}$-map).

Among the configurations that are compared, the UVAM results in the best performance with nearly 2.7× increase in execution speed compared to the case without visual attention. The execution speed is directly related to the keypoint reduction factor γ and attention overhead τ as described by Eq. (7). The low keypoint reduction factor ($\gamma = 0.18$) of the UVAM overweighs the negative effects of its relatively high attention overhead. Both the bottom-up saliency only case and top-down familiarity only case suffer from relatively high keypoint reduction factor owing to their low attention accuracy. The average recognition rate for each of the attention configurations is 95% with no false positive matches.

Recognition rate must be kept equal for each of the attention configurations in order for execution time to be meaningful as a performance metric. In our test setup, the recognition accuracy is

Fig. 11. Objects and background images used for test image generation. (a) Since we are interested in the performance of the attention model, a subset of 75 of the more easily recognizable objects were chosen from the COIL-100 object database to reduce the impact of the limitations of SIFT based object recognition. (b) Background images are categorized into three groups according to the amount of salient clutter they contain.

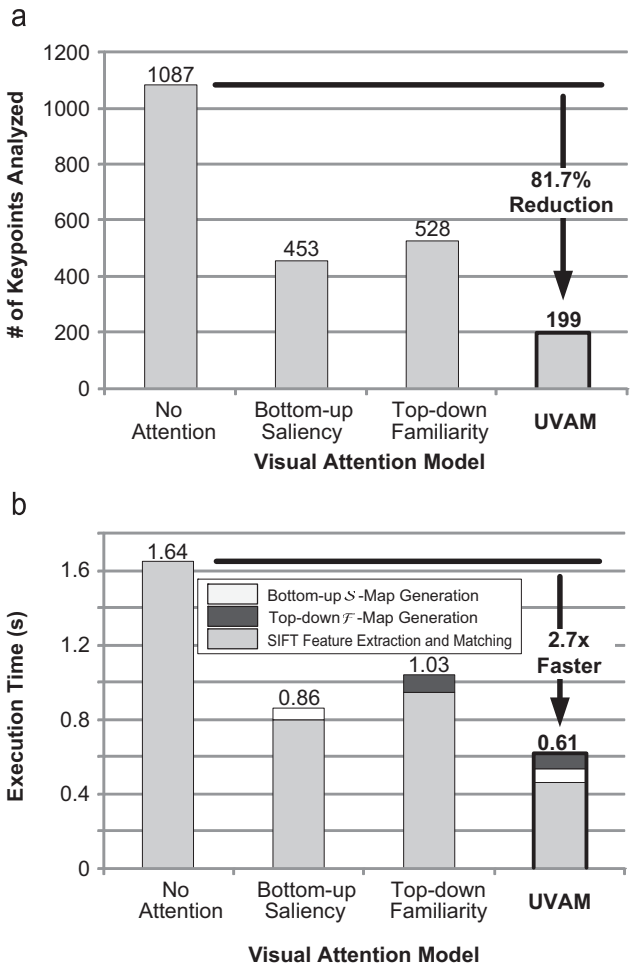

Fig. 12. Performance summary of the proposed UVAM compared to different configurations of visual attention. (a) The number of analyzed keypoints, and (b) the execution times of each configuration are compared.

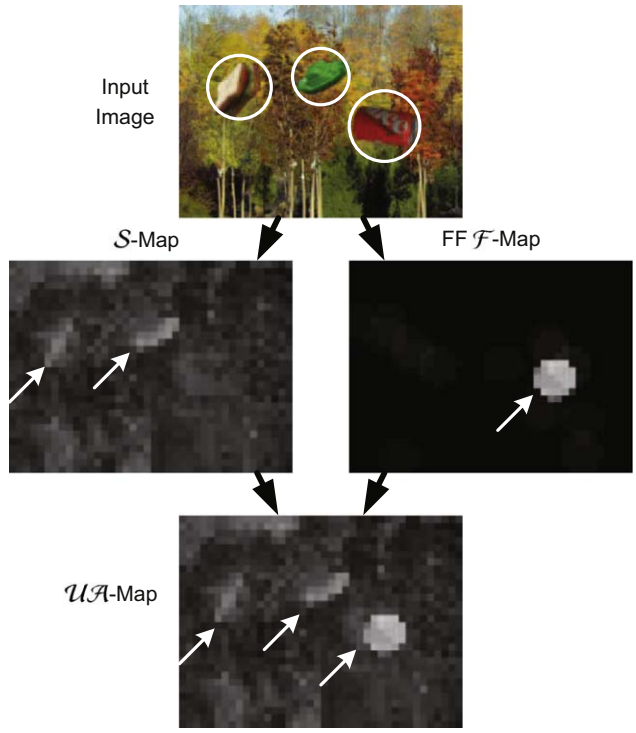

Fig. 13. Complementary operation of the bottom-up $\mathcal{S}$-map and the top-down $\mathcal{F}$-map.

determined solely by the underlying SIFT recognition since the attention models keep selecting ROIs either until all objects are recognized or the entire scene is analyzed. This means that all objects that are recognizable by SIFT are eventually recognized by each of the attention configurations. Only the number of visited ROI, and thus the execution time, will vary from configuration to configuration.

It should be noted that Walther's attention based recognition system [7,8], which also uses SIFT, employs a different test method to show that visual attention can actually improve recognition rate. In his experiment, the number of allowed attention fixations is limited to 5, thereby effectively keeping the execution time constant. As a result, the recognition rate is constantly higher when visual attention is used, compared to when random fixation is used. While this method successfully illustrates the benefits of attention, it is not as suited as a practical object recognition system, since the recognition rate is actually lower than what is possible using SIFT alone due to the limited number of allowed fixations.

4.2. Robustness to target object type

The high efficiency of the UVAM stems from the complementary nature of its bottom-up and top-down parts. The $\mathcal{S}$-map and the $\mathcal{F}$-map respond more strongly to different but complementary types of objects, thus increasing the chance that target objects get attention. In addition, the two different attentions are vulnerable to two distinct types of background clutter, making it unlikely for both attention mechanisms to fail at once.

The bottom-up and top-down mechanisms of visual attention are suited for detecting different types of objects. The bottom-up mechanism is most effective at detecting objects that consist of non-textured solid surfaces. This is because the $\mathcal{S}$-map promotes intensity, color, or orientation features that occur as single peaks in the feature map. Objects that have a lot of detailed textures tend to produce multiple peaks instead of a single strong peak, and are inhibited due to competition.

Top-down attention is most effective for objects that are large and heavily textured as they generally produce more keypoints with large scale than small non-textured objects. This is because the input image resolution is reduced by a factor of λ (in this case 0.5) prior to preliminary object recognition for the FF $\mathcal{F}$-map generation. This reduction in resolution effectively filters out keypoints of smaller scale. Increasing object size has the effect of increasing the scale of its keypoints, thus improving the chances of those keypoints being detected during the FF $\mathcal{F}$-map generation stage. Meanwhile, for objects of the same size, textured objects produce more keypoints than objects consisting of smooth surfaces. For the COIL-100 objects used for test image generation, the number of keypoints extracted ranges from 6 to 94 depending on their texture content. As a result, recognition performance is greatly dependent on the target objects.

Fig. 13 clearly shows the complementary operations of the top-down and bottom-up attentions. The $\mathcal{F}$-map shows a strong response for the textured soda can but misses the other two objects. The $\mathcal{S}$-map, on the other hand, responds strongly to the two objects missed by the $\mathcal{F}$-map. When the two are combined into the unified attention map, all three objects are correctly detected as shown in the bottom of Fig. 13.

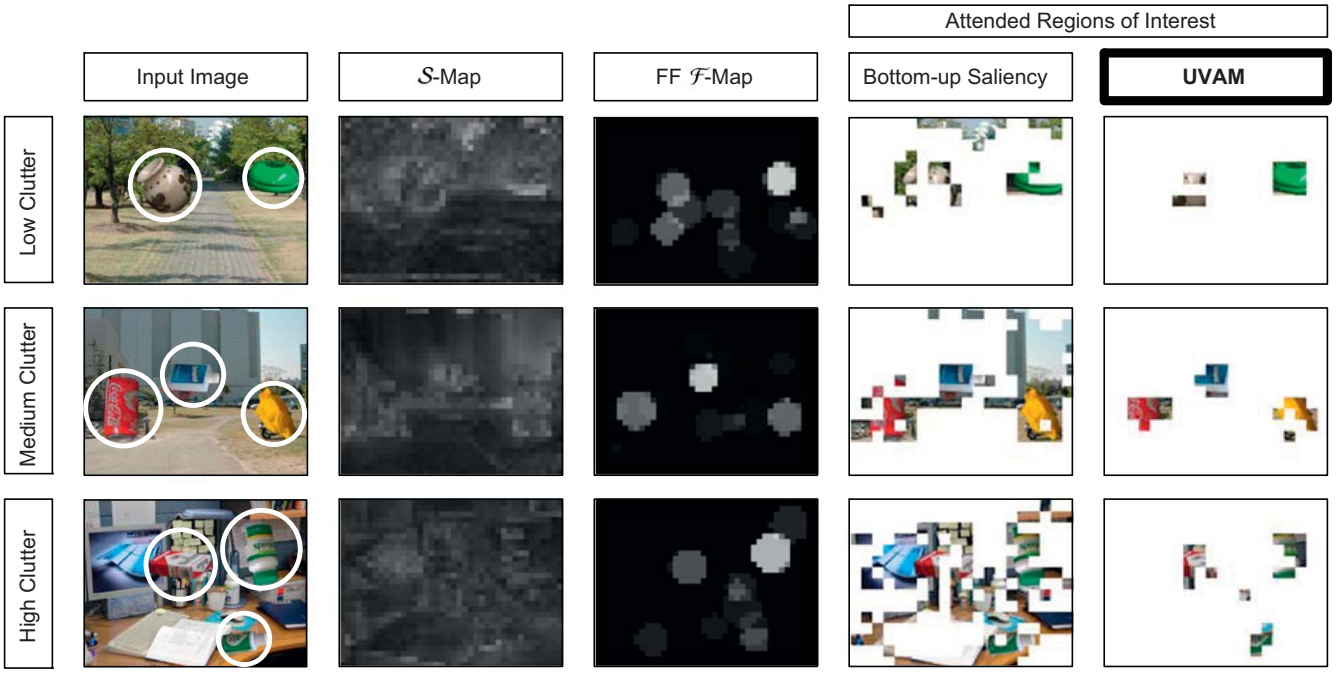

Fig. 14. Comparison of the performance of bottom-up saliency based visual attention, and the UVAM for three scenes with varying amounts of salient clutter. The number of attended ROIs increases proportionally to the amount of salient clutter for bottom-up saliency based visual attention, while it does not increase substantially for the proposed UVAM.

4.3. Robustness to background clutter

According to Fig. 12, the bottom-up only visual attention case needs to analyze more keypoints than the UVAM. This is because it selects inaccurate ROI in scenes with large amounts of "salient" clutter. The amount of salient clutter in a scene can be quantified as the percentage of the image corresponding to background clutter that has saliency exceeding the saliency value of the least salient target object. Fig. 14 shows the ROI selection results for three scenes with low (15%), medium (27%), and high (42%) salient clutter with only bottom-up visual attention compared to those with the UVAM. As the amount of salient clutter increases, the $\mathcal{S}$-map becomes less representative of the locations of the target objects in the scene. This results in a greater number of ROIs, and thus keypoints, being attended to before all trained objects are recognized.

Bottom-up visual attention is prone to salient clutter due to its method to generate the $\mathcal{S}$-map [4]. As outlined in Section 2.1, the $\mathcal{S}$-map is generated through the combination of intensity, color, and orientation features that stand out from its surroundings. In the $\mathcal{S}$-map, features of the same type must compete with each other for attention. For example, while blue and red features are generated by separate feature extraction processes, they are eventually combined into a single color feature map. As a result, even a single feature in the background that is salient in terms of its intensity, color, or orientation may inhibit the responses for all features of the same type.

Meanwhile, the performance of top-down attention is not affected by salient background clutter but can be adversely affected by "familiar" clutter, which is a totally different type of clutter arising from distractors that exhibit high familiarity. While it has been shown in Section 2.2 that distractor keypoints originating from non-targets have low probability of exhibiting high familiarity, occasionally the net familiarity of many distractor keypoints concentrated in a region may overwhelm the familiarity of target keypoints.

Salient clutter and familiar clutter are not highly correlated as can be seen by comparing the $\mathcal{S}$-maps and FF $\mathcal{F}$-maps in Fig. 14. Therefore when the $\mathcal{S}$-map and $\mathcal{F}$-map are combined into the $\mathcal{UA}$-map as proposed in this paper, only regions that correspond to target objects are reinforced, making the $\mathcal{UA}$-map very robust to background clutter.

4.4. Failure mode of the UVAM

The UVAM may fail to correctly predict locations of target objects under certain conditions. The most common failure mode is when the $\mathcal{S}$-map fails due to salient clutter and the $\mathcal{F}$-map fails due to objects that are either too small or do not contain enough texture or both. While failure of the $\mathcal{S}$-map is solely dependent on the input image, failure of the $\mathcal{F}$-map can be alleviated by increasing the input scaling factor λ and decreasing matching error ε, as explained in Section 2.2. This, however, requires more computational power and results in increased execution time of visual attention.

Another cause of failure for the visual attention model lies in the limitation of the reference object recognition system itself. As previously pointed out, the recognition rate is 95% regardless of the configuration of the visual attention. For images containing the 5% of objects which are not successfully recognized, the ROI selection process continues until the entire image is attended to, leading to increased execution time.

4.5. Robustness on natural images

Further tests are carried out on two sets of natural images to evaluate the robustness of the UVAM and confirm the results obtained using the synthesized images in the previous subsections. The first test set, which was used in [31], contains 51 test images composed of eight objects. The second test set, which was photographed for this study, contains 75 test images composed of 10 objects. All test

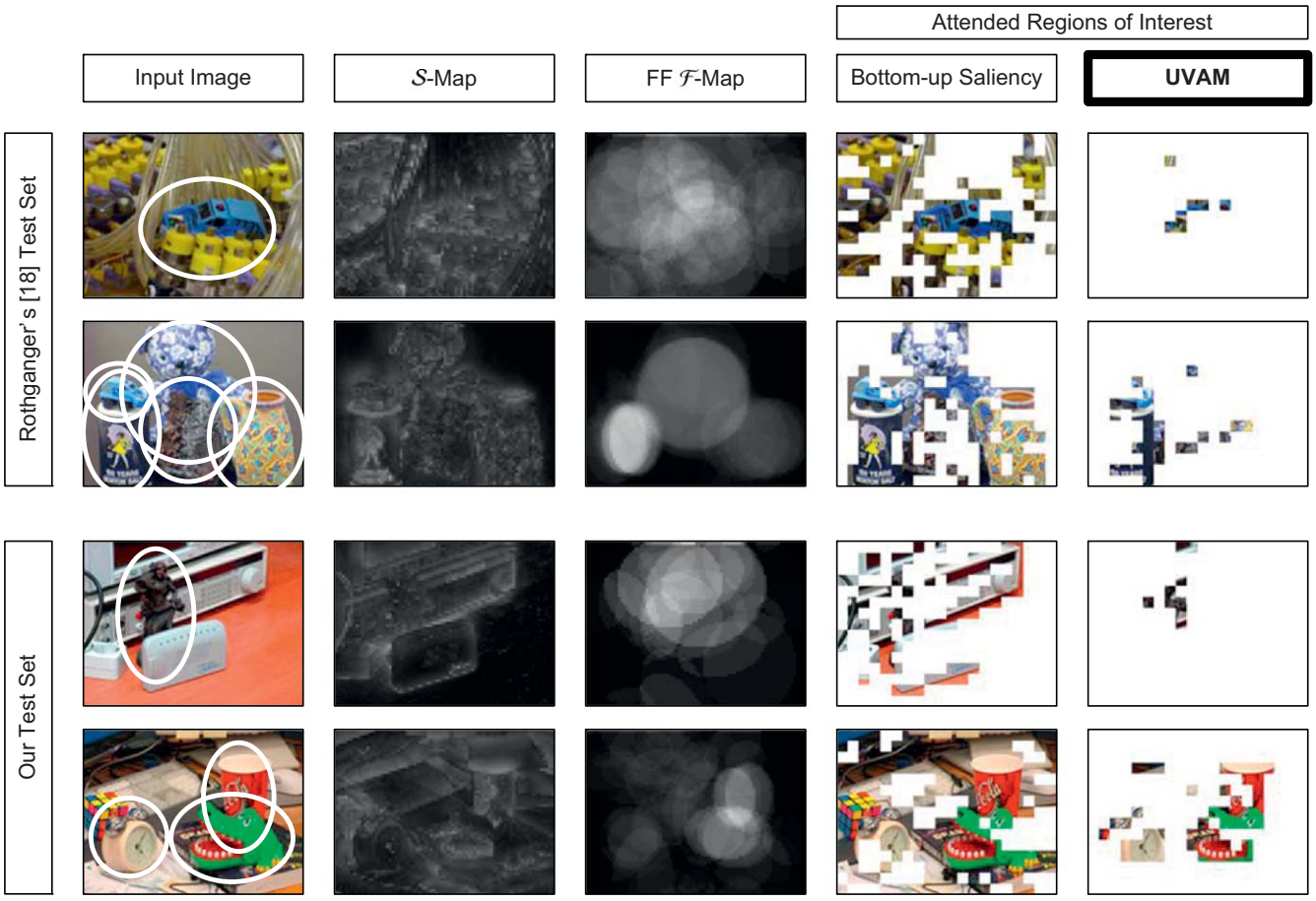

Fig. 15. Testing of the UVAM on natural images. The results are similar to those for the synthesized test images.

Table 1
Performance of unified attention model on natural images.

Test set	Attention mode	Recognition rate (%)	ROIs	Keypoints	Execution time (s)
Rothganger [20]	No attention	91	300	1494	3.91
	Bottom-up saliency	91	55	491	2.28
	UVAM	91	43	332	1.92
Our images	No attention	94.2	300	1972	3.01
	Bottom-up saliency	94.2	56	548	1.81
	UVAM	94.2	24	229	1.44

images are scaled to 1280×960 pixels resolution, which is the resolution used in [31]. The ROI size is accordingly increased to 64×64 pixels to maintain a constant total ROI count. For both test sets the keypoint database is constructed using images of each object taken from different views varying by approximately 30° increments.

Fig. 15 shows ROI selection results for the UVAM compared to bottom-up attention. The performance of our model on the natural images is summarized in Table 1. Recognition rates for both test sets are above 90% regardless of the visual attention configuration with no false positives. The 91% recognition rate achieved for the test images of Rothganger et al. [31] is comparable to that of the various methods that were compared in that paper. On average more than 2× gains in recognition speed are obtained for both test sets with the UVAM applied.

5. Conclusion

This paper proposes the unified visual attention model (UVAM), which combines stimulus-driven bottom-up attention and goal-driven top-down attention to reduce execution time of object recognition. The SIFT object recognition flow is analyzed, and the UVAM is integrated to reduce the number of analyzed keypoints with optimizations to minimize attention overhead. The UVAM is quantitatively evaluated using 3600 synthesized images, with further testing on 126 natural images to check for robustness.

The main contribution of the UVAM is its use of familiarity as a top-down component of attention. By using familiarity to guide attention towards known objects, visual attention performance is substantially improved compared to when only bottom-up saliency is adopted. Also, since familiarity is calculated using SIFT features, many computations can be shared with the object recognition flow and the overhead of attention can be minimized. Applying the UVAM model to object recognition of the synthesized images resulted in 2.7× speed-up without reduction in recognition accuracy. Further tests on the natural images resulted in around 2× speed-up without reduction in recognition accuracy. These results show that the UVAM is an effective and robust model of visual attention for speeding up SIFT based object recognition systems.

References

[1] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision, 1999, pp. 1150–1157.
[2] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[3] U. Neisser, Cognitive Psychology, Appleton-Century-Crofts, New York, 1967.
[4] J.K. Tsotsos, The complexity of perceptual search tasks, in: International Joint Conferences on Artificial Intelligence, 1989, pp. 1571–1577.
[5] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 124–1259.
[6] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Human Neurobiology 4 (1985) 219–297 916.
[7] D. Walther, U. Rutishauser, C. Koch, P. Perona, Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, Computer Vision and Image Understanding 100 (1–2) (2005) 41–63.
[8] U. Rutishauser, D. Walther, C. Koch, P. Perona, Is bottom-up attention useful for object recognition?, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2 (2004) 37–44.
[9] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (2006) 1395–1407.
[10] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[11] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, et al., Curious George: an attentive semantic robot, Robotics and Autonomous Systems, 2008.
[12] R. Desimone, J. Duncan, Neural mechanisms of selective visual attention, Annual Review of Neuroscience 18 (1995) 193–222.
[13] D. Walther, L. Fei-Fei, Task-set switching with natural scenes: measuring the cost of deploying top-down attention, Journal of Vision 7 (11) (2007) 1–12.
[14] J.H. Fecteau, D.P. Munoz, Salience, relevance, and firing: a priority map for target selection, Trends in Cognitive Sciences 10 (8) (2006) 382–390.
[15] J.J. Bonaiuto, L. Itti, Combining attention and recognition for rapid scene analysis, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Workshops, 2005.
[16] V. Navalpakkam, L. Itti, An integrated model of top-down and bottom-up attention for optimizing detection speed, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2 (2006) 2049–2056.
[17] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artificial Intelligence 78 (1995) 507–545.
[18] A. Olivia, A. Torralba, M.S. Castelhano, J.M. Henderson, Top-down control of visual attention in object detection, IEEE International Conference on Image Processing 1 (2003) 253–256.
[19] G. Deco, B. Schürmann, A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition, Vision Research 40 (2000) 2845–2859.
[20] D.C. Donderi, D. Zelnicker, Parallel processing in visual same-different decisions, Perception and Psychophysics 5 (1969) 197–200.
[21] J. Christie, R.M. Klein, Familiarity and attention: does what we know affect what we notice?, Memory and Cognition 23 (1995) 547–550.
[22] D. Soto, D. Heinke, G.W. Humphreys, Early, involuntary top-down guidance of attention from working memory, Journal of Experimental Psychology: Human Perception and Performance 31 (2) (2005) 248–261.
[23] J.K. Tsotsos, Y. Liu, J.C. Martinez-Trujillo, M. Pomplun, E. Simine, K. Zhou, Attending to visual motion, Computer Vision and Image Understanding 100 (2005) 3–40.
[24] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, T. Poggio, Robust object recognition with cortex-like mechanisms, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 411–426.
[25] R.L. Sproull, Refinements to nearest-neighbor searching, Algorithmica 6 (1991) 579–589.
[26] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, A.Y. Wu, An optimal algorithm for approximate nearest neighbor searching in fixed dimensions, Journal of the ACM 45 (6) (1998) 891–923.
[27] S.A. Nene, S.K. Nayar, H. Murase, Columbia object image library (Coil-100), Technical Report CUCS-006-96, Columbia University, February 1996.
[28] D.H. Ballard, Generalizing the Hough transform to detect arbitrary shapes, Pattern Recognition 13 (2) (1981) 111–122.
[29] M. Brown, D.G. Lowe, Invariant features from interest point groups, in: Proceedings of British Machine Vision Conference, 2002, pp. 656–665.
[30] L.J. Heyer, S. Kruglyak, S. Yooseph, Exploring expression data: identification and analysis of coexpressed genes, Genome Research 9 (11) (1999) 1106–1115.
[31] F. Rothganger, S. Lazebnik, C. Schmid, J. Ponce, 3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints, International Journal of Computer Vision 66 (3) (2006) 231–259.

About the Author—SEUNGJIN LEE received the B.S. and M.S. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2006 and 2008, respectively, and is currently working toward the Ph.D. degree in electrical engineering and computer science at KAIST. He has been involved with the development of digital hearing aids and vision processors. Currently, his research interests include bio-inspired computer vision algorithms and heterogeneous multi-core architectures for computer vision SoCs.

About the Author—KWANHO KIM received the B.S. and M.S. degrees in electrical engineering and computer science from Korea Advanced Institute of Science and Technology (KAIST) in 2004 and 2006, respectively. He is currently working toward the Ph.D. degree in electrical engineering and computer science at KAIST. In 2004, he joined the Semiconductor System Laboratory (SSL) at KAIST as a Research Assistant. His research interests include VLSI design for object recognition, architecture and implementation of NoC-based SoC.

About the Author—JOO-YOUNG KIM received the B.S. and M.S. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2005 and 2007, respectively, and is currently working toward the Ph.D. degree in electrical engineering and computer science at KAIST. Since 2006, He has been involved with the development of the vision processors. Currently, his research interests include bio-inspired vision algorithm and parallel architecture for computer vision system.

About the Author—MINSU KIM received the B.S. and M.S. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea in 2007 and 2009, respectively. He is currently working toward the M.S. degree in electrical engineering and computer science at KAIST. His research interests include network-on-chip based SoC design and VLSI architecture for computer vision processing.

About the Author—HOI-JUN YOO graduated from the Electronic Department of Seoul National University, Seoul, Korea, in 1983 and received the M.S. and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 1985 and 1988, respectively. His Ph.D. work concerned the fabrication process for GaAs vertical optoelectronic integrated circuits.
From 1988 to 1990, he was with Bell Communications Research, Red Bank, NJ, where he invented the 2D phase-locked VCSEL array, the front-surface-emitting laser, and the high-speed lateral HBT. In 1991, he became a manager of the DRAM design group at Hyundai Electronics and designed a family of fast-1 M DRAMs to 256 M synchronous DRAMs. In 1998, he joined the faculty of the Department of Electrical Engineering at KAIST and now is a full professor. From 2001 to 2005, he was the director of System Integration and IP Authoring Research Center (SIPAC), funded by Korean Government to promote worldwide IP authoring and its SoC application. From 2003 to 2005, he was the full time Advisor to Minister of Korea Ministry of Information and Communication and National Project Manager for SoC and Computer. In 2007, he founded System Design Innovation and Application Research Center (SDIA) at KAIST to research and to develop SoCs for intelligent robots, wearable computers and bio systems. His current interests are high-speed and low-power network on chips, 3D graphics, body area networks, biomedical devices and circuits, and memory circuits and systems. He is the author of the books *DRAM Design* (Seoul, Korea: Hongleung, 1996; in Korean), *High Performance DRAM* (Seoul, Korea: Sigma, 1999; in Korean), *Low-power NoC for High-performance SoC Design* (CRC Press, 2008), and chapters of *Networks on Chips* (New York, Morgan Kaufmann, 2006).
Dr. Yoo received the Electronic Industrial Association of Korea Award for his contribution to DRAM technology in 1994, the Hynix Development Award in 1995, the Design Award of ASP-DAC in 2001, the Korea Semiconductor Industry Association Award in 2002, the KAIST Best Research Award in 2007, and the Asian Solid-State Circuits Conference (A-SSCC) Outstanding Design Awards in 2005, 2006 and 2007. He is an IEEE fellow and serving as an Executive Committee Member and the Far East Secretary for IEEE ISSCC, and a Steering Committee Member of IEEE A-SSCC. He was the Technical Program Committee Chair of A-SSCC 2008.