Robust Face Detection Based on Knowledge-Directed Specification of Bottom-Up Saliency

Yu-Bu Lee and Sukhan Lee

This paper presents a novel approach to face detection by localizing faces as the goal-specific saliencies in a scene, using the framework of selective visual attention of a human with a particular goal in mind. The proposed approach aims at achieving human-like robustness as well as efficiency in face detection under large scene variations. The key is to establish how the specific knowledge relevant to the goal interacts with the bottom-up process of external visual stimuli for saliency detection. We propose a direct incorporation of the goal-related knowledge into the specification and/or modification of the internal process of a general bottom-up saliency detection framework. More specifically, prior knowledge of the human face, such as its size, skin color, and shape, is directly set to the window size and color signature for computing the center of difference, as well as to modify the importance weight, as a means of transforming into a goal-specific saliency detection. The experimental evaluation shows that the proposed method reaches a detection rate of 93.4% with a false positive rate of 7.1%, indicating the robustness against a wide variation of scale and rotation.

Keywords: Face detection, goal-specific visual attention, knowledge-directed specification, bottom-up saliency.

Manuscript received Mar. 15, 2010; revised Nov. 3, 2010; accepted Dec. 21, 2010.

This work was partially supported by WCU program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2010-000-10062-0), by PRCP through NRF of Korea funded by MEST (2011-0018397), and by the KORUS-Tech program (KT-2010-SW-AP-FS0-0004).

Yu-Bu Lee (phone: +82 31 299 6487, email: basilia@skku.edu) is with the School of Information and Communication Engineering, Sungkyunkwan University, Suwon, Rep. of Korea.

Sukhan Lee (email: lsh@ece.skku.ac.kr) is with the Department of Interaction Science and School of Information and Communication Engineering, Sungkyunkwan University, Suwon, Rep. of Korea.

doi:10.4218/etrij.11.1510.0123

I. Introduction

Vision systems that have the task of operating in real-world environments must tackle challenges that arise from the specific conditions and the uncertainties found in the visual information of the captured images [1]. In particular, when searching for an object in real-world scenes, selective attention is necessary to focus on the information that is relevant to a current task and to efficiently process the appropriate data sources. Attention is the process of selecting and gating visual information based not only on the saliency found in the image itself (bottom-up) but also in prior knowledge about scenes, objects, and their interrelations (top-down) [2]. Bottom-up attention directs the gaze to salient regions obtained from features based on the basic information of an input image, such as intensity, color, and orientation. Top-down attention determines the salient locations through perceptive processing, such as understanding and recognition.

Several computational models have been proposed to simulate a human's visual attention using a bottom-up computational framework [3]-[8]. Itti and others [3] proposed a bottom-up model and built a system called Neuromorphic Vision C++ Toolkit. After that, Walther [6] extended this model to address proto-object regions and created the Saliency Toolbox (STB). He also applied this model to accomplish object recognition tasks [7]. However, the high computational cost and the limited choice of parameters were weaknesses found in these models. Gao and others [9] presented a discriminant saliency detection model which requires a discriminant saliency selection process at the training stage. The saliency map can then be computed by the selected features at the testing stage. In current machine attention methods, bottom-up selection plays an important role in

providing early cues in a multistage competitive attention processing scheme [10]. For object detection, the bottom-up selective attention method is often used as a preprocessing step to reduce the demand on the image processing capacity, that is, the detection of salient regions and a restriction of the search region to only the necessary fractions of the input image.

Most traditional object detection models use a sliding window across the image and apply a binary classifier at each window to detect the desired target object [11]. While this approach has been successfully applied to the detection of rigid objects, such as faces, cars, and pedestrians [11]-[15], it is slow and computationally expensive since each classifier is run independently at every window within the image. To overcome this speed bottleneck, object detection models use an attention operator to rapidly select a few points of interest in the image. However, most such models use either a purely top-down or bottom-up approach.

There have been few attempts to integrate both the top-down and bottom-up attention [8], [16]-[19]. Navalpakkam and Itti [16] enhance the bottom-up salience model to yield a simple yet powerful architecture to learn the target objects from the training images containing targets in diverse and complex backgrounds. Ramström and Christensen [17] calculate the feature and background statistics to be used in a game theoretic winner-takes-all (WTA) framework for the detection of objects. Choi and others [18] suggest the learning of the desired modulations of the saliency map for a top-down tuning of attention with the aid of an ART network. Frintrop proposed the VOCUS model [8], which consists of bottom-up and top-down maps. The bottom-up part is based on Itti's model, while the top-down part uses previously learned weights to enable the search for targets. The weighted features contribute to a top-down saliency map, highlighting the regions with the target relevant features. The total saliency map is a linear-combination of the two maps using a fixed user-provided weight. Such integration models are useful for object detection in many applications, such as human computer interaction, human robot interaction (HRI), robot navigation, visual surveillance, and any realistic visual search. Notably, face detection is a crucial task in HRI, which is necessary for natural interactions between a human and a robot. Lee and others [19] showed that the interactive spiking neural network (ISNN) can be used to bias the bottom-up processing for face detection. They use skin color, facial features, and an ellipse-like shape for determining a bottom-up map that is correlated with the cued features. The network can then manipulate the amount of bottom-up and top-down influences on a search task needed to investigate the dynamic and modulatory aspects of the selective attention. Ban and others [20] suggested a face detection model that integrates a bottom-up mechanism for extracting features and a top-down perceptual mechanism for perceiving facial features, such as the face form and color. They construct a face conspicuity map by binding the bottom-up process and the top-down process consisting of the face form and color feature maps. The main focus of visual attention is on how to integrate the bottom-up and top-down information in order to obtain an efficient decision on where to focus the attention within the input image [1]. In previous models [19], [20], a single face saliency map was constructed by integrating the bottom-up and top-down maps generated using the features extracted from the original image. For drawing a rapid focus of attention on a target, an integration mechanism needs to direct the bottom-up process to a specific saliency detection process through the top-down feedback using the knowledge relevant to the target.

In this paper, a novel approach to goal-specific visual attention, such as detecting the target object as saliency, is presented. The approach is based on transforming a general purpose of bottom-up saliency detection process into a specific optimal saliency detection process for finding the target object. The transformation to a goal-specific saliency detection process is done by incorporating the knowledge on the goal or the target object directly into the specification and/or modification of the bottom-up process. The bottom-up process is configured to be general yet flexible enough to accommodate such modification and/or specification. Specifically, it can be constructed by revising a conventional bottom-up saliency detection process as in [3], [8], or the combination of both, in such a way as to be able to carry out the knowledge-based direct modification and/or specification.

Some instances of the knowledge-based direct modification and/or specification of a bottom-up process follow. i) The known color signature of the target object can be fed to the bottom-up color saliency detection process to set the reference for computing the center-surround difference. ii) The known shape of the target object can also set the references for computing the center-surround difference for the respective bottom-up shape saliency detection processes. iii) They can help determine the weights to be assigned to the saliencies generated by other features at each layer of the Gaussian pyramid in such a way as to signify the particular shape known for the target object. iv) The known size of the target object, together with the 3D depth information available, can determine the size of the reference window to be used for the computation of the center-surround difference as well as the weights to be assigned to the saliencies detected at each layer of the Gaussian pyramid. The proposed approach differs from the conventional approaches to the integration of the bottom-up and top-down saliency detection [8], [19], [20] in that it uses the knowledge relevant to the target to directly encode the

bottom-up framework of saliency extraction, instead of either by training the summation weights based on the performance of the target object classification [8] or by generating the top-down saliency map through evaluating of the saliencies defined by the bottom-up map [19], [20]. The proposed model uses 3D depth information as well as the 2D information, such as size, shape, and skin color distribution, as the knowledge related to the face for rebuilding a general structure of the bottom-up process into the human face saliency detection process. Our model successfully detected the face regions against scale and rotation variations by providing guidance to likely face locations through the multiplicative top-down weights on the bottom-up feature maps.

The rest of this paper is organized as follows. In section II, the structure of our model is briefly introduced, and then the bottom-up module for constructing the feature maps based on skin color in detail is discussed. This is then followed by a description of the top-down processing method used to obtain the task-relevant knowledge. Section III presents the experimental results and performance evaluations. Finally, we draw some brief conclusions in section IV.

II. Goal-Specific Visual Attention Model

The proposed goal-specific visual attention model integrates both the bottom-up and top-down influences in order to guide the attention by the knowledge related to the face during a face finding visual search. Figure 1 illustrates the schematic diagram of the proposed goal-specific visual attention model for face detection. As shown in Fig. 1, the proposed model consists of a bottom-up module and top-down module, and the face saliency is determined by integrating the bottom-up saliency and top-down knowledge.

The proposed bottom-up module adopts the bottom-up part of the VOCUS proposed by Frintrop [8]. In the saliency-based bottom-up module, we use the skin color feature for extracting a face-like area and a center-surround Haar-like feature to compute the center-surround difference. Finally, we construct three feature maps by summing up the two scale maps found within each scale. The feature maps are based on the scale invariant saliency, which makes it possible to detect the salient regions at an image scale matching the face size.

The top-down module can help drive the bottom-up processing to its goal of face detection by providing feedback to the processing stages of the bottom-up module, as shown in Fig. 1. In order to draw attention to the regions with face characteristics, the proposed top-down module generates knowledge relevant to the face, such as depth, face size corresponding to distance, face shape, and skin color distribution, as prior knowledge built into the human mind. We describe the details of each module in the following subsections.

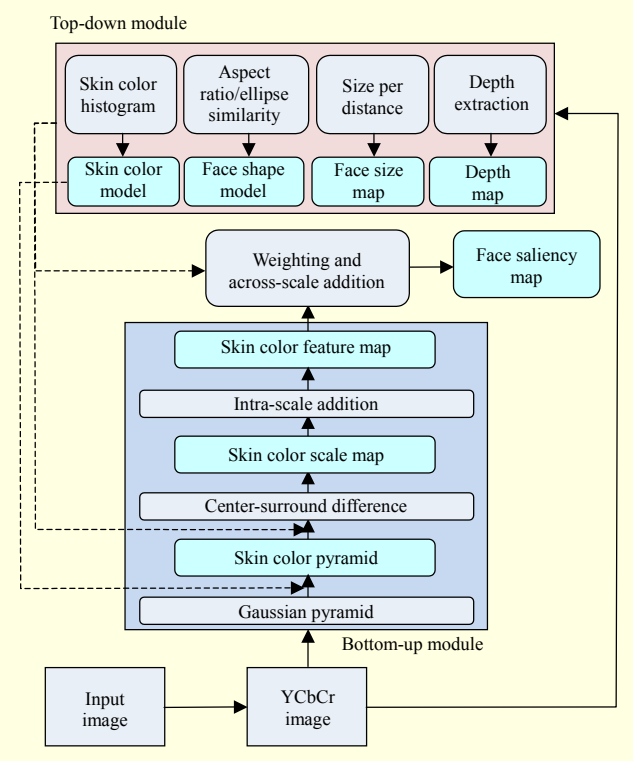


Fig. 1. Architecture of proposed visual attention model.

1. Bottom-Up Module

A. Skin Color Pyramid

For feature computation in our bottom-up saliency-based module, we only use a skin color feature instead of the luminance, orientation, and color features used in VOCUS since the skin color feature easily enables the candidate face areas to be located in color images.

First, we convert the color input image into the YCbCr image. The YCbCr color space is usually employed for video storage and coding and can provide an effective analysis for human skin color [21]. From the YCbCr image, a Gaussian image pyramid is computed by applying a 3×3 Gaussian filter to the image. Then, the image is subsampled. This results in an image that is half the width and height of the original one. This process is repeated four times, resulting in an image pyramid with five different scales, s_0 to s_4. Finally, from the Gaussian image pyramid, we generate a skin color pyramid P by

$$P_s(x,y) = \exp\left\{-\left(\frac{(Cr(x,y)-\mu_{Cr})^2}{2\sigma_{Cr}^2} + \frac{(Cb(x,y)-\mu_{Cb})^2}{2\sigma_{Cb}^2}\right)\right\}, \quad (1)$$

where $P_s(x, y)$ is the pixel value in the skin color pyramid at

(a) (b)

Fig. 2. Skin color pyramid with five different scales 0 to 4: (a) input image and (b) skin color pyramid P_0 to P_4.

scale s; $Cr(x, y)$ and $Cb(x, y)$ are the Cr and Cb values at pixel (x, y), respectively; and μ and σ denote the mean and the standard deviation of the skin color distribution for the Cr and Cb component, respectively. We employ $\mu_{Cr} = 150$, $\sigma_{Cr} = 15$, $\mu_{Cb} = 105$, and $\sigma_{Cb} = 25$, which are determined by the histogram analysis of the skin color distribution computed in the top-down module. Figure 2 shows the results of the skin color pyramid with five different scales, 0 to 4. The following computations are all performed on scales 2 to 4. This makes the model robust to noise since no noise pixels occur in the smoothed images [8].

B. Scale Maps

We create the skin color scale maps by calculating the center-surround differences in the three skin color pyramid images that represent scales 2, 3, and 4. To compute the center-surround operation in our model, we modify the structure of the center-surround differences used in VOCUS. VOCUS represents the center c with one pixel within each scale and determines the surround by computing the average of the surrounding pixels for the two different surround sizes with a radius σ of 3 or 7 pixels, respectively. Finally, the center-surround operation calculates the difference between the center value and the surround value. Unlike VOCUS, we define the center c as a rectangular region, not as a pixel, for computing the center-surround differences. Using a rectangular region as the center, it is possible to roughly detect not only the region with a similar skin color but also the region which is well fitted to the face size. For the rectangular region, we use the center-surround Haar-like feature as shown in Fig. 3. The center-surround Haar-like feature is one of the Haar-like features proposed by Viola and Jones [14] for face detection.

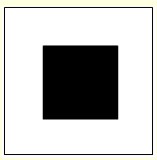


Fig. 3. Center-surround Haar-like feature used for computing center-surround difference.

(a) (b) (c)

Fig. 4. Skin color scale maps of scale 2: (a) input image, (b) scale map by center-surround Haar-like feature with $(\sigma_c, \sigma_s) = (3, 7)$, and (c) scale map by center-surround Haar-like feature with $(\sigma_c, \sigma_s) = (7, 10)$.

The center-surround Haar-like feature can be represented as a black and a white rectangle. Its output is the difference between the sums of the pixel values within the two rectangular regions. However, we compute the difference between the averages of the pixel values within two regions (the center region and surround region) using the center-surround Haar-like feature as the reference window as follows:

$$center(x, y, s, \sigma_c) = \frac{\sum_{x'=-\sigma_c}^{x'=\sigma_c} \sum_{y'=-\sigma_c}^{y'=\sigma_c} P_s(x+x', y+y')}{(2\sigma_c+1)^2}, \quad (2)$$

$$surround(x, y, s, \sigma_s) = \frac{\sum_{x'=-\sigma_s}^{x'=\sigma_s} \sum_{y'=-\sigma_s}^{y'=\sigma_s} P_s(x+x', y+y') - center(x,y,s,\sigma_c)}{(2\sigma_s+1)^2 - (2\sigma_c+1)^2}, \quad (3)$$

$$ScaleMap_{s,\sigma_c,\sigma_s}(x, y) = \max\{center(x, y, s, \sigma_c) - surround(x, y, s, \sigma_s), 0\}, \quad (4)$$

where s represents the image scale with $s \in \{2, 3, 4\}$; and σ_c and σ_s denote the radii of the center region and the surround region, respectively. In our model, we use two different sizes of the center-surround feature with $(\sigma_c, \sigma_s) = (3, 7)$ and $(\sigma_c, \sigma_s) = (7, 10)$, respectively. Both center region sizes are determined by information of the size of the face taken from the face size map constructed in the top-down module. $ScaleMap_{s,\sigma_c,\sigma_s}(x, y)$ denotes the computation of the center-surround difference of pixel (x, y) with (σ_c, σ_s) in the scale s image. Using (2) through (4), we obtain six scale maps. Figure 4 illustrates the two skin-color scale maps generated using the two different sizes of the center-surround feature in the skin color pyramid of scale 2.

C. Feature Maps

After computing the scale maps, we construct three feature maps using intra-scale addition. In VOCUS, the feature map is generated by across-scale addition in which all the scale maps are resized to a larger scale and then added up pixel by pixel. In

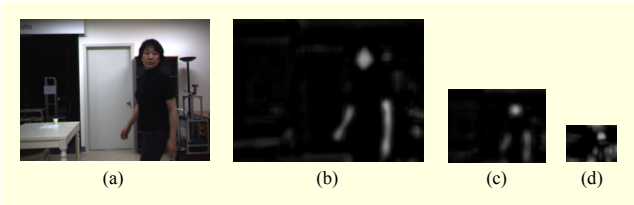


Fig. 5. Skin color feature maps generated by intra-addition within each scale: (a) input image, (b) feature map of scale 2, (c) feature map of scale 3, and (d) feature map of scale 4.

our model, we construct three feature maps using intra-scale addition instead of the across-scale addition used in [8]. In other words, we sum up pixel by pixel the two scale maps from each scale. As a result, this yields three different scales of the skin color feature maps.

The feature map is based on scale invariance, which enables the features to stand out at an image scale matched to their sizes. For instance, large features will be highlighted at the coarse scale. Also, small features will be highlighted at the fine scale [25]. In addition, by intra-scale addition, we can construct feature maps with various resolutions, just like the human vision system. The different scales of the three feature maps can reflect the results from various view distances occurring between the object and camera. If a human face is at a long distance from the camera, the size of the face region is proper using the coarser resolution. The different scales of the feature maps are used for discriminating the face area matching the size of the face that corresponds to the distance from the candidate face area for the weighting processing. Figure 5 shows the three feature maps generated by adding up the two scale maps from within each scale.

2. Top-Down Module

A. Skin Color Model Based on Histogram Analysis

Skin color is a very effective and strong cue for finding faces since the hue of skin is roughly invariant across different ethnic groups, and skin color distribution lies in a limited chromatic color space [19]. To build the skin color model, we compute the skin color distribution for the *Y*, *Cr*, and *Cb* components using a histogram analysis. In order to investigate skin color distribution, we manually segmented the faces in 600 test images containing the human face and then performed the histogram analysis for the statistical information of the chrominance and the luminance. The histogram results for chrominance *Cr* and *Cb* components are distinctly formed into a Gaussian-similar distribution rather than into a uniform distribution. The empirical ranges of the *Cb* and *Cr* components for the skin color are typically Cb_{skin} = (77, 127) and Cr_{skin} = (133, 173) [21]. In our model, we use Cb_{skin} = [80, 130], Cr_{skin} = [135, 165], and Y_{skin} = [60, 120], derived from our histogram results. The values of the chrominance for facial skin color are narrowly distributed indeed, while the luminance values are not at all narrowly distributed in the face area. Different density values can be found distinctly at the interval (0, 255) [26]. Nevertheless, we use the luminance component *Y* of the skin color since it is useful to discriminate between the skin regions and the non-skin regions with the chrominance range used for skin color.

The histogram results of the skin color chrominance for the *Cr* and *Cb* components and the luminance for the *Y* component are used for constructing the skin color pyramid in the bottom-up module and for weighting the salient regions in the face feature maps, respectively.

B. Face Shape Model Based on Aspect Ratio and Ellipse Similarity

Whereas skin color is an effective feature for face detection, it also detects other body parts, for example, naked hands and legs, as well as skin-color noise. Thus, two face features, the aspect ratio and the ellipse similarity, are introduced to discriminate face region from skin color regions.

Aspect ratio. The aspect ratio between the width of the face and the height of the face is a unique face cue that discriminates it from other body parts. The golden ratio of an ideal face, calculated by Govindaraju [27], is height/width = $(1+\sqrt{5}/2)$. For computing the aspect ratio, we segment the face region on the test images and then obtain its principal directions from the segmented face region by applying a principal component analysis (PCA). We calculate the covariance matrix for the *X* and *Y* coordinates of the approximate region and then determine the major and minor directions by calculating the eigenvectors and the eigenvalues of the covariance matrix. As a result, we can compute the aspect ratio of the face from the ratio between the length of the minor and major axes which correspond to the width and the height of the face, respectively. For our shape model, we define 1.5 as the value of the face aspect ratio using empirical results.

Ellipse similarity. The shape of the human face takes the form of an ellipse. In our face shape model, we apply an ellipse similarity method to search for a face shape. We measure the ellipse similarity of each segmented blob using the major and minor axes as follows:

$$E = \frac{4 \times R}{\pi \times l_1 \times l_2}, \tag{5}$$

where *R* is the number of pixels in the segmented blob; and l_1 and l_2 denote the lengths of the minor and major axes, respectively. With respect to the resultant value of (5), if the

ellipse similarity value is 1, the segmented blob is considered to be a face region.

C. Face Size and Depth Map Based on Depth Information

Stereo information obtained from a stereo vision system is used to extract objects from the scene. In the top-down module, we generate a depth map by using depth information obtained from the stereo vision system. The depth map represents the existence of one or more solid objects at a particular depth. This depth map is applied for its depth information of a skin color segmented region in the bottom-up module.

In addition to the depth map, we produce a face size map by estimating the face size corresponding to the distance for identifying the face area. The face size map is used to consider whether the size of each region extracted from the skin color is the proper size that corresponds to its distance. In other words, if the size of an extracted region is a reasonable size for a face region corresponding to the distance by referring to the face size map, it has high probability of being a face. This information is very helpful in discriminating between a face region and a non-face region. To construct a face size map, we first captured images containing the human face at certain distances in the range of 0.5 m to 4 m at 0.5 m intervals. Next, from the obtained test images, we measured the approximate size of the face in three different resolutions, 160×120, 80×60, and 40×30, which are then applied for the three different scales of the skin color feature maps. The face size corresponding to all the spatial distances is calculated by applying a polynomial fitting to the empirical results. Finally, we construct a face size map corresponding to the distance for the three different resolutions. The face size map performs a role as the reference map representing the face size corresponding to the depth obtained from the stereo camera. Thus, the face size map makes it possible to discriminate between a face and a non-face region by providing information for the reasonable size of a face region.

3. Integrating Bottom-Up Saliency and Top-Down Knowledge

In our model, the top-down knowledge obtained from multiple cues, such as depth, size, shape, and luminance, and the bottom-up saliency obtained from the skin color feature are integrated through a weighting processing stage, which produces a single face saliency map. By integrating the bottom-up and top-down knowledge, each candidate region within the three feature maps is weighted using the weighting functions based on the top-down information, and then the weighted feature maps are summed up by across-scale addition as follows:

$$\begin{aligned} SM &= \bigoplus_{s} W(F_s), \\ W(F_s) &= \sum_i W^i_{\text{size}} \cdot F^i_s + W^i_{\text{shape}} \cdot F^i_s + W^i_{\text{intensity}} \cdot F^i_s, \\ i &\in \{0, 1, \ldots, k\}, \end{aligned} \quad (6)$$

where s represents the scale with $s \in \{2, 3, 4\}$; $\oplus$ denotes the across-scale addition; i is a candidate region segmented using the connected component analysis at each feature map; and $W(F_s)$ denotes the weighted feature map for each scale s. For computing the face probability of each candidate region, region i is weighted using the three weighting functions as follows:

$$W^i_{\text{size}} = \exp\left\{-\left(-\frac{(i_{\text{size}} - \mu_{\text{size}})^2}{2\sigma^2_{\text{size}}}\right)\right\}, \quad (7)$$

$$W^i_{\text{shape}} = \exp\left\{\left(-\frac{(i_{\text{ratio}} - \mu_{\text{ratio}})^2}{2\sigma^2_{\text{ratio}}}\right) + \left(-\frac{(i_{\text{ellipse}} - \mu_{\text{ellipse}})^2}{2\sigma^2_{\text{ellipse}}}\right)\right\}, \quad (8)$$

$$W^i_{\text{luminance}} = \exp\left\{\left(-\frac{(i_{\text{luminance}} - \mu_{\text{luminance}})^2}{2\sigma^2_{\text{luminance}}}\right)\right\}, \quad (9)$$

where i_{size}, i_{ratio}, i_{ellipse}, and $i_{\text{luminance}}$ denote the size, aspect ratio, ellipse similarity, and average of the luminance value of the candidate region i, respectively, and (μ_{size}, σ_{size}), (μ_{ratio}, σ_{ratio}), (μ_{ellipse}, σ_{ellipse}), and ($\mu_{\text{luminance}}$, $\sigma_{\text{luminance}}$) are the mean and standard deviation of the face size according to the depth, aspect ratio, ellipse similarity, and luminance, respectively. The estimates of the mean and standard deviation are defined from the size map, shape model, and skin color model built in the top-down module. According to (7) through (9), each candidate region outputs a probability value of the face. That is, the proposed model filters out every candidate blob within all the feature maps using the weighting functions. The three feature maps weighted by combining bottom-up with top-down modules are summed up to a single face saliency map. In a face saliency map, the face regions are represented by the more salient areas than the non-face regions since they have higher face probability.

To detect the face regions in the face saliency map, we first create a binary image used for finding the salient areas above a certain threshold, and then perform a connected component analysis. We determined that 30% of the maximum value is proper for the saliency threshold value. The optimal threshold value was determined through a tradeoff between the detection ratio and the false positive ratio using the empirical analysis. We show the empirical analysis results for determining the threshold in section III. To visualize the face region detected, we use a rectangle determined by the width and height of the salient area.

III. Experimental Results

For our experiments, we applied our method to 471 images obtained from a stereo vision camera which captures 15 frames per second with a 640×480 resolution. The test images contain single and multiple faces with variations in scale, orientation, and pose from diverse indoor environments. All our evaluations were performed on an Intel Pentium Core 2 PC with a 2.53 GHz CPU and 2 GB of main memory.

To evaluate the performance of the proposed method, we have tested three main experiments: (i) the detection of faces under different conditions, (ii) a comparison to the Viola and Jones face detector, which is a commonly used technique for face detection, and (iii) a comparison with the VOCUS and STB models, which are visual attention models based on different visual features. Figure 6 shows the results of face detection for distances in the range of 0.5 m to 4 m. Figures 6(a) and (b) depict the face saliency map using the proposed model and the detected face using face saliency map, respectively. Figure 7 demonstrates that the proposed model can detect frontal and non-frontal single faces under different poses and backgrounds. As shown in Fig. 7, our model can successfully detect a non-frontal face with rotation variations (pan rotation) in the range of –90º to 90º as well as the frontal face located in cluttered background.

Figure 8 illustrates that our model detected multiple faces under different poses and distances. Each face has different face views and also different scales due to their position to the camera. Figure 8 shows the detection of highly rotated faces in pan (first row), roll (second row), and tilt (third row) rotations at different positions from the camera. From these results, we can see that our model is able to correctly detect face regions regardless of scale and rotation variations because our method considers not only the face size according to the distance from the camera for scale invariance but also the luminance distribution of the face region and the shape, such as in the aspect ratio and ellipse similarity using the PCA for rotation invariance.

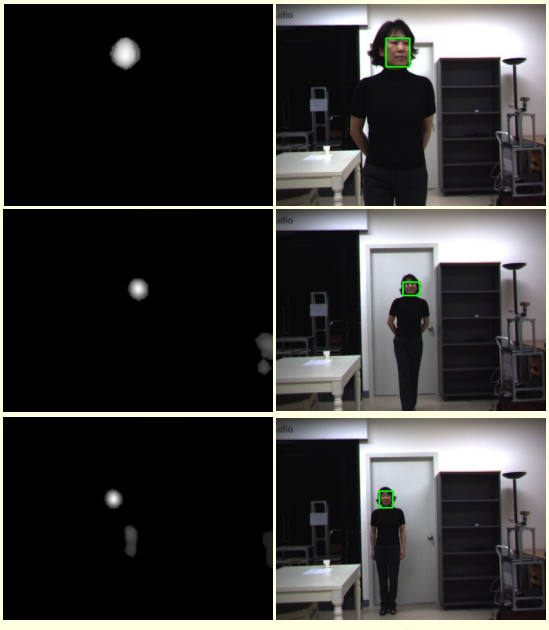
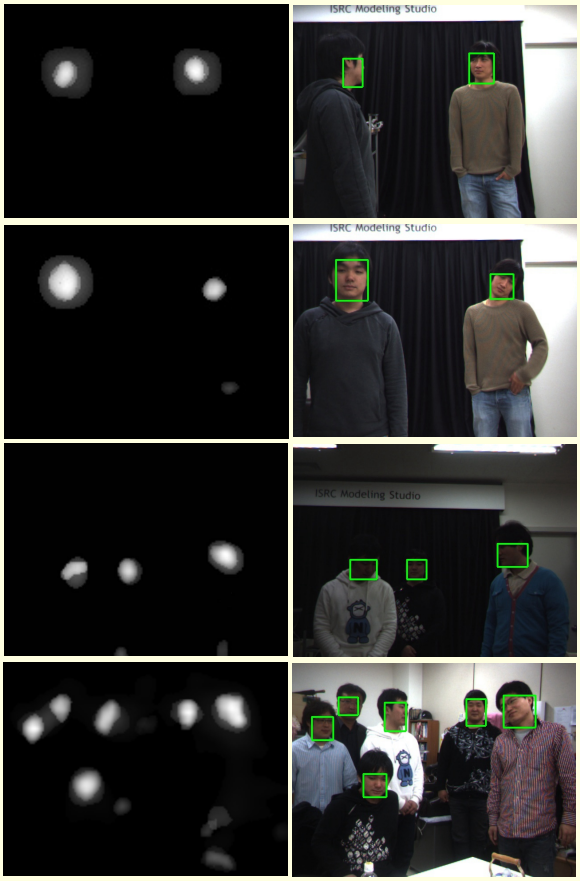


Fig. 6. Face detection results for different distances: (a) face saliency map using proposed model and (b) detected face from (a).

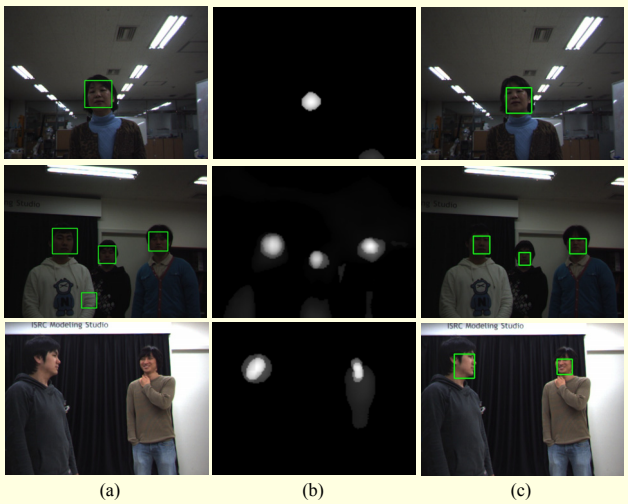
Fig. 7. Face detection results for single face under different poses and backgrounds: (a) face saliency map using proposed model and (b) detected face from (a).

In Fig. 9, we compared the proposed model with the Viola and Jones face detector, which is provided by the OpenCV implementation [22]. We used the training data from OpenCV to detect the frontal [23] and profile [24] human faces. The Viola and Jones face detector is a fast face detection system based on an extended set of Haar-like features and an Ada-Boost classifier. This system is commonly used for face detection due to its real time operation and accuracy. By comparing Figs. 9(a) and (c), we can see that our model produces more correct detections. In Fig. 9(a), we can see the results with false (second row) and missing (third row) detections by the Viola and Jones face detector, while our model accurately detects faces without false and missing detections. We believe that this is because our model correctly detects faces with scale and rotation variations by considering multiple cues obtained from the top-down knowledge.



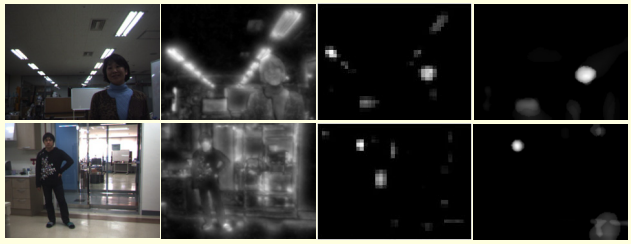
(a) (b)

Fig. 8. Face detection results for multiple faces under different poses: (a) face saliency map using proposed model and (b) detected faces from (a).



(a) (b) (c)

Fig. 9. Comparisons of face detection: (a) face detection results by Viola and Jones face detector, (b) face saliency maps by proposed model, and (c) face detection results from (b).



(a) (b) (c) (d)

Fig. 10. Comparisons between saliency maps of our model and previous visual attention models: (a) original image and saliency maps generated (b) by bottom-up part of VOCUS, (c) by STB, and (d) by proposed model.

We also compared the saliency map between the previous visual attention models and the proposed model. In our comparisons, we used the bottom-up part of VOCUS using the three features of intensity, orientation, and color and STB (latest version available at http://www.saliencytoolbox.net) by adding on skin hue to the feature list. The saliency map comparisons of the proposed model and the previous models on the selected images are shown in Fig. 10. It is evident that the proposed model outperforms the others. In Fig. 10(d), our method detects the salient face regions correctly, while the VOCUS detects non-face regions as salient regions because it does not take the features relevant to the target into account as shown in Fig. 10(b). Similarly, in Fig. 10(c), the STB also produced false detection results in spite of considering skin color as a feature.

In our experiments, a face is successfully detected if the local image of the search window contains the eye and mouth. Otherwise, it is a false positive. The detection rate is the ratio between the number of successful detections and the total number of faces in the test images. The false positive rate is the ratio between the number of false positives and the total number of searching windows.

In our method, for determining an optimal threshold value used in the face saliency map, we analyzed 400 test images. Figure 11 illustrates the plotted graphs for the detection rate/false positive rate versus the threshold. As shown in Fig. 11, the detection rate and false positive rate were the lowest when the threshold is 0.9, whereas they were the highest when the threshold is 0.6. We can see both the detection rate and the false positive rate decrease when the threshold value increases. Therefore, we determined that 0.7 (corresponding to 30% of the maximum value in the face saliency map) would be used as a threshold value for providing both a high detection rate and a low false positive rate at the same time.

Table 1 summarizes the detection results for seven different combinations of cues, which are size (size/depth), shape, and luminance of skin color distribution. These cues have been

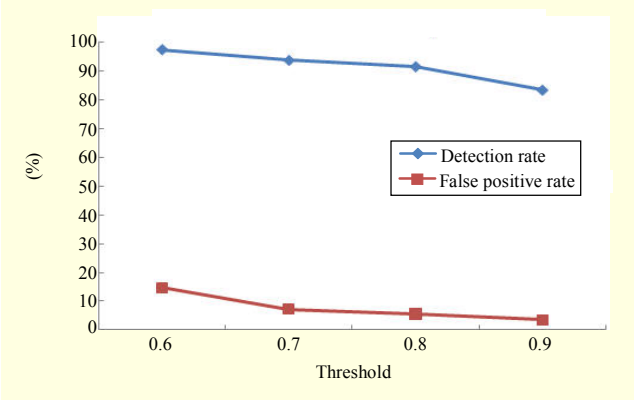


Fig. 11. Plots of detection rate/false positive rate vs. threshold.

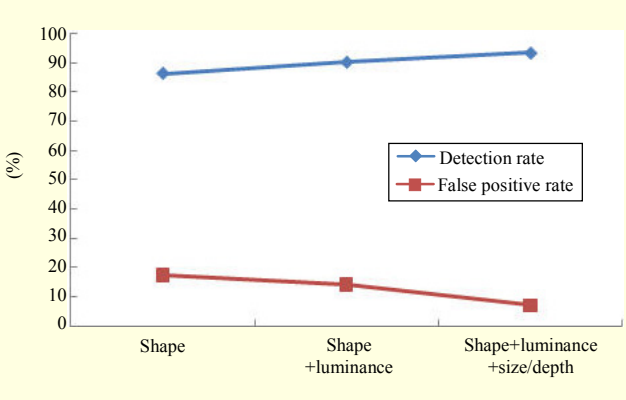


Fig. 12. Performance plots for combination of cues.

Table 1. Face detection results using different combinations of cues on 471 test images.

Cues enabled			Detection rate (%)	False positive rate (%)
Shape	Luminance	Size/depth		
√			86.3	17.4
	√		84.4	27.1
		√	88.9	18.3
√	√		90.2	14.1
√		√	91.7	11.5
	√	√	91.0	14.3
√	√	√	93.4	7.1

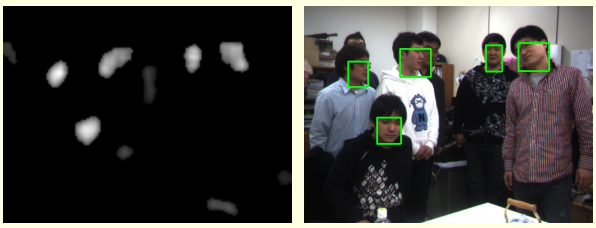


Fig. 13. Example of missing detections due to occlusion: (a) face saliency map using our model and (b) face detection results from (a).

Table 2. Comparison results between the Viola and Jones face detector and proposed model on 150 frontal face images.

Method	Detection rate (%)	False positives
Viola and Jones	91.3	23
Proposed	97.6	8

used as weighting constraints for determining the likelihood of a skin region being a face region. To evaluate the performance of each cue, we have tested with the various combinations of cues on 471 test images. In Table 1, a size cue is denoted as size/depth since depth information is used to compute if the size of skin color region is the appropriate face size according to the range of depth. In the case of a single cue, its false positive rate is high. In particular, the power of a luminance cue is relatively low as compared to a shape or size/depth cue even if the luminance is still useful for eliminating the non-skin regions with the chrominance range of skin color. As shown in Table 1, the shape or size/depth cue provides important effects of reducing the number of false positives. Especially, combining size/depth and shape (or luminance) provides a higher detection rate and a lower false positive rate than the combination of both shape and luminance because 3D information compensates the intrinsic weakness of a 2D cue by estimating the possible face size from depth information. As the number of cues increases, the false positive rate is decreased drastically (up to 20%), while preserving a high detection rate. In Table 1, our results indicate a detection rate of 93.4% and a false positive rate of 7.1% by using the multiple cues. For a distinct comparison, Fig. 12 illustrates the performance plots for the combination of cues. In Fig. 12, we observe that the proposed method can improve the performance by a combination of three cues, that is, shape + luminance + size/depth.

Figure 13 shows an example of missing detections due to a partial occlusion and multiple faces of close proximity. As shown in Fig. 13(a), overlapped faces that are closely located result in a salient region from merging connected skin parts.

Table 2 presents the comparison results between the proposed model and the Viola and Jones face detector on 150 images including frontal faces under a wide variation of backgrounds. From the information point of view, the Viola and Jones face detector is based on 2D information obtained from monocular image, while our method is an integrative approach of 2D and 3D information which is obtained from a

stereo camera. The key purpose of our comparison is to show how much improvement the 3D information may incur for face detection, indicating the degree the proposed model can improve the detection accuracy by combining 2D and 3D information. In Table 2, we find that our model provides higher detection rates and less false positives than the Viola and Jones face detector. This is in part due to the fact that the proposed method uses real-world measurements based on stereo vision as well as shape and luminance cues in order to filter out non-face regions.

IV. Conclusion

In this paper, we have proposed a goal-specific visual attention model based on knowledge-directed specification of bottom-up saliency for detecting human face. The proposed approach transforms a general purpose of bottom-up saliency detection process into a goal-specific saliency detection process for detecting the human face. For transforming to a goal-specific saliency detection process, the top-down module directly encodes the bottom-up framework by using 3D depth information as well as 2D cues, such as size, shape, and skin color distribution, as the knowledge relevant to the face. Through the top-down influence based on knowledge-directed specification of bottom-up processing, our model correctly detects faces against scale and rotation variations since it considers not only the appropriate face size from depth information for scale invariance but also the luminance distribution of the skin color and shape, such as the aspect ratio and ellipse similarity using PCA for the rotation invariance. Experimental results show that our model accurately detects faces under various poses. Our results indicate a detection rate of 93.4% and a false positive rate of 7.1%. However, our model still has both false positives due to noisy non-face areas detected from illumination variations and missing detections due to occlusion.

In future work, we plan to improve the skin color model to increase the robustness of the face detection against illumination variations. In addition, we will utilize the model for an HRI application. Therefore, the runtime (currently 0.31 s on a 2.53 GHz Pentium Core for 640×480 pixel images) has to be improved to enable real-time performance.

References

[1] L. Paletta, E. Rome, and H. Buxton, "Attention Architectures for Machine Vision and Mobile Robots," *Neurobiology of Attention*, New York: Academic Press, 2005, pp. 642-648.

[2] L. Itti and C. Koch, "Computational Modeling of Visual Attention," *Nat. Rev. Neurosci.*, vol. 2, no. 3, 2001, pp. 194-203.

[3] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, 1998, pp. 1254-1259.

[4] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vis. Res.*, vol. 40, 2000, pp. 1489-1506.

[5] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Proc. NIPS*, 2005, pp. 547-554.

[6] D. Walther and C. Koch, "Modeling Attention to Salient Proto-objects," *Neural Netw.*, vol. 19, 2006, pp. 1395-1407.

[7] D. Walther et al., "Attentional Selection for Object Recognition–A Gentle Way," *Lect. Notes Comput. Sci.*, vol. 2525, no. 1, 2002, pp. 472-479.

[8] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal Directed Search*, PhD thesis, University of Bonn, Germany, 2005.

[9] D. Gao, V. Mahadevan, and N. Vasconcelos, "The Discriminant Centersurround Hypothesis for Bottom-Up Saliency," *NIPS*, 2007.

[10] V. Navalpakkam and L. Itti, "An Integrated Model of Top-Down and Bottom-Up Attention for Optimizing Detection Speed," *Proc. IEEE CVPR*, vol. 2, 2006, pp. 2049-2056.

[11] C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection," *Int. J. Comput. Vision*, vol. 38, no. 1, 2000, pp. 15-33.

[12] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network Based Face Detection," *IEEE Trans. PAMI*, vol. 20, no. 1, 1998, pp. 23-38.

[13] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," *Proc. IEEE CVPR*, 2000, pp. 746-751.

[14] P. Viola and M.J. Jones, "Robust Real-Time Face Detection," *Int. J. Computer Vision*, vol. 57, no. 2, 2004, pp. 137-154.

[15] P. Viola, M.J. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," *Int. J. Comput. Vision*, vol. 63, no. 2, 2005, pp. 153-161.

[16] V. Navalpakkam and L. Itti, "Sharing Resources: Buy Attention, Get Recognition," *Proc. Int. Workshop Attention Performance Comput. Vision*, 2003.

[17] O. Ramström and H. Christensen, "Object Detection Using Background Context," *Proc. Int. Conf. Pattern Recognition*, 2004, pp. 45-48.

[18] S. Choi, S. Ban, and M. Lee, "Biologically Motivated Visual Attention System Using Bottom-Up Saliency Map and Top Down Inhibition," *Neural Info. Process.-Lett. Review*, vol. 2, no. 1, 2004, pp. 19-25.

[19] K. Lee, H. Buxton, and J. Feng, "Selective Attention for Cue Guided Search Using a Spiking Neural Network," *Proc. Workshop Attention Performance Comput. Vision*, 2003, pp. 55-63.

[20] S.W. Ban, M. Lee, and H.S. Yang, "A Face Detection Using

Biologically Motivated Bottom-Up Saliency Map Model and Top-Down Perception Model," *Neuro Comput.*, vol. 56, 2004, pp. 475- 480.

[21] N. Habili, C.C. Lim, and A. Moini, "Segmentation of the Face and Hands in Sign Language Video Sequences Using Color and Motion Cues," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 14, no. 8, 2004, pp. 1086-1097.

[22] Intel Corporation. OpenCV: Open Source Computer Vision Library, http://www.intel.com/research/mrl/research/opencv/

[23] R. Lienhart and J. Maydt, "An Extended Set of Haar-Like Features for Rapid Object Detection," *IEEE Conf. Image Process.*, vol. 1, 2002, pp. 900-903.

[24] D. Bradley, "Profile Face Detection," Intel Research Award Contest, 2003.

[25] L. Feng and G. Michael, "Region Enhanced Scale-Invariant Saliency Detection," *IEEE Int. Conf. Multimedia Expo*, 2006, pp. 1477-1480.

[26] H. Li and K.N. Ngan, "Saliency Model Based Face Segmentation and Tracking in Head-and-Shoulder Video Sequences," *J. Visual Commun. Image Representation*, vol. 19, no. 5, 2008, pp. 320-333.

[27] V. Govindaraju, "Locating Human Faces in Photographs," *Int. J. Comput. Vision*, vol. 19, no. 2, 1996, pp. 129-146.

Yu-Bu Lee received her BS, MS, and PhD in computer science and engineering from Ewha Womans University, Seoul, Rep. of Korea, in 1990, 1992, and 2008, respectively. From 2008 to 2009, she was a post doctor in the Intelligent Systems Research Center, Sungkyunkwan University, Suwon, Rep. of Korea. Since 2010, she has been a research professor in the School of Information and Communication Engineering, Sungkyunkwan University. Her current research interests include visual attention, human-robot interaction, cognitive robotics, and medical image processing and analysis.

Sukhan Lee received his BS and MS in electrical engineering from Seoul National University in 1972 and 1974, respectively, and the PhD in electrical engineering from Purdue University, West Lafayette, in 1982. From 1983 to 1997, he was a professor with the Department of Electrical Engineering and Computer Science, University of Southern California. From 1990 to 1997, he worked as senior member of technical staff for the Jet Propulsion Laboratory, NASA/California Institute of Technology. From 1998 to 2003, he served as an executive vice president and chief research officer of the Samsung Advanced Institute of Technology. Since 2003, he has been a professor of information and communication engineering, WCU professor of interaction science, and the director of the Intelligent Systems Research Center at Sungkyunkwan University, Suwon, Rep. of Korea. He is a fellow of IEEE and of Korea Academy of Science and Technology. His research interests include cognitive robotics and vision, intelligent systems, and micro/nano electro-mechanical systems.