I J C T A, 9(41) 2016, pp. 591-598 ISSN: 0974-5572

ATTENDING SALIENT FACE IN THE CROWD USING TOP DOWN FOLLOWED BY BOTTOM UP MECHANISM

Ravi Kant Kumar [*], Jogendra Garain [*], Dakshina Ranjan Kisku [*] and Goutam Sanyal[*]

Abstract: Our visual system attending a scene having human faces, differently, than a normal scene. While attending a scene containing faces, we normally ignore the other non-face things. It is human physiology and our social behavior nature that intend us to give more importance to the human faces rather than other objects. Therefore, during visiting a scene, at first level of view we just simply attend the faces available in the images and ignore the other objects or background. Next, at finer level of view, reasoning and decision takes place in the deeper level of brain and our focus goes towards the most prominent face in the crowd. Therefore, in order to mapping these concepts in the same way, top down and bottom up attention mechanisms has been incorporate for the computer vision system. Deeper level attention improves the face recognition and identification evolution in a crowd image with complex objects and background. Hence this innovative idea may boost up the visual surveillance, security and robotic vision tasks. Various computer vision-based techniques have been carried out to determine selective attention and saliency of objects in a scene. But, in the context of finding salient faces, hardly some researchers have been explored. This paper contributes a novel idea to analyze and find out the salient faces in the crowd.

Key Words: *Face Attention, Visual Surveillance, Visual Security, Top down Saliency, Bottom up Saliency, Face Detection.*

1. INTRODUCTION

We have an immense ability of attending the high-flying regions in a visual scene. Humans constantly collect information and intermingle with their surrounds things, via five senses [1]. This sensory and visual information goes to the deeper level of the brain. But due to limited processing capacity of our brain, it analyses, interprets and intelligently filters the unwanted visual information. Therefore, for detection recognition and identification [2] only the selected information goes to the deeper level of the brain. This phenomenon is called selective attention [3]. It is directed by a mechanism called visual saliency [4, 5].

In the context of human visual system, when we keep our target to see human faces in a scene, we are selectively approachable to some stimuli (faces) above others (non face objects). This mechanism is called top down [6, 7] approach for attending a scene. But among these selected stimuli (faces), exactly on which face our gaze will focus, is decided by the 'dominating features' carrying out in that face. The face wearing 'dominating features' is considered as salient face.

[*] **Department of Computer Science and Engineering, National Institute of Technology, Durgapur, India**
vit.ravikant@gmail.com, jogs.cse@gmail.com, drkisku@gmail.com, nitgsanyal@gmail.com

Attending a face based on the 'dominating features' of its stimulus is a bottom up approach [6, 7]. In general, visual attention limits our focus towards the 'standalone region' in that visual scene. Visual attention is one of the most imperative incident through that we intelligently performs tasks like scene recognition, Object Tracking [8], Robot Navigation [9], Intelligent camera having saliency estimator [10], Automatic target detection [11] etc. Till date it is not well understandable about all the aspects and factors that guide our attention. However, two renowned mechanisms are mainly responsible for the attention task, are; top-down and bottom-up approaches.

Human vision system cannot handle a large amount of visual information at a time, therefore, it filter and process only the prominent or desired information from the scene. Computer vision researchers got motivated by this human way of attending a scene and somehow manage to develop various visual attention models. The aim of these models is to provide visual ability of the computer system in the similar way to human vision system and to reduce the processing time and cost by grabbing only the desired or salient locations of the scene. Some of these models are; Feature integration theory [12] that tells about several visual search approaches, is the base of many of the researches conceded in successive decades. Koch and Ullman [3] modeled a fundamental framework using feature-integration theory [12] and bottom-up mechanism [6, 7]. Moreover, Center-Surround Difference Based Model [5], composite saliency indicator (CSI) [13], graph-based model in [14] has also been explored later.

Inclusions of visual saliency in the area of face detection and surveillance, so far, little research work have been done. Some of them are, Attention capture by faces [15], and Visual perception based on eye movements [16], Selective attention-based method for face recognition [17], 3D face recognition using Kinect [18], Saliency map augmentation with facial detection [19] and Context Aware Saliency[20] etc.

This paper recommends a novel framework for visiting the salient face in the crowd. In a scene, many salient regions other than faces may also available. Therefore, the first part of the work is target specific. Here, the target is to determine the face region in the scene voluntary. In second part of work is to find the salient face among the detected faces, based on the low level features of the stimulus of faces. Hence, this framework includes the combination of top down and bottom up approaches.

Rest of this paper is organized as follows: In Section II, bottom up and top down mechanism for visual attention in the context of face recognition is discussed. Framework of Proposed and existing system has been discussed in Section III. The proposed methodology has been briefly explained in Section IV. Experimental result and discussion is presented in Section V. Finally, Section VI appeals the concluding notes and future work.

2. VISUAL ATTENTION APPROACHES

a) Bottom Up Approach

Bottom up approach works based on task independent mechanism for guiding user attention uses towards the salient objects in the visual area by focusing our gaze to the regions having high saliency. Bottom up mechanism [6, 7], accord with the attractiveness of stimulus (face) with respect to its surrounding faces in that scene. This attractiveness occurs due to some of the prominent features in terms of color, intensity and orientation. It determines the popping out of a particular face from a scene at a premature stage of our vision without any prior acquaintance or knowledge about the scene. Therefore, bottom up mechanism is an unguided, signal-based exogenous technique. This approach is purely driven by exterior stimuli.

b) Top Down Approach

Top down approach works based on a specific task/goal to direct our attention. At the time of searching for a particular object in a scene, the top down mechanism extract the dominating features from that object and focus our gaze towards the region where that object belonging in the visual field. We recognize and acquainted with our family, relatives or known person in a mob, very promptly and proficiently because, these identified faces give extra attention to us. This takes place as we previously have some priory information of that known person. Therefore, our recognition system acquires an instant response about the known person from the brain and which direct our eye to focus on that face. The attending and recognizing the faces in the set of multiple faces (crowd) in this fashion is known as the top down attention [6, 7].

In this paper, saliency system has been framed after combining these two mechanisms (Top down and Bottom up). Modeling computer vision system along this way (By combining the top town and bottom up approaches) with the inclusion of more visual parameters, a robust and efficient model for visual attention can be developed in the task of face detection and recognition.

3. PROPOSED SYSTEM

Bottom up and Top down mechanism of visual attention works together independently. The proposed method applied the concept of and Top Down followed by Bottom Up approach for finding the salient face location in the scene. The first work of our model is target specific i.e. selection of faces and to ignore the non-face regions even then non-face regions may have the more attentive locations. Therefore, Top down mechanism has been applied to detect the faces in the given images. Next, work is to finding the most salient face among the detected faces. For this, saliency has been estimated purely based on stimulus information of the faces. Bottom up mechanism has been applied for it.

For better understanding, framework for general saliency model as well as proposed saliency model has been described below.

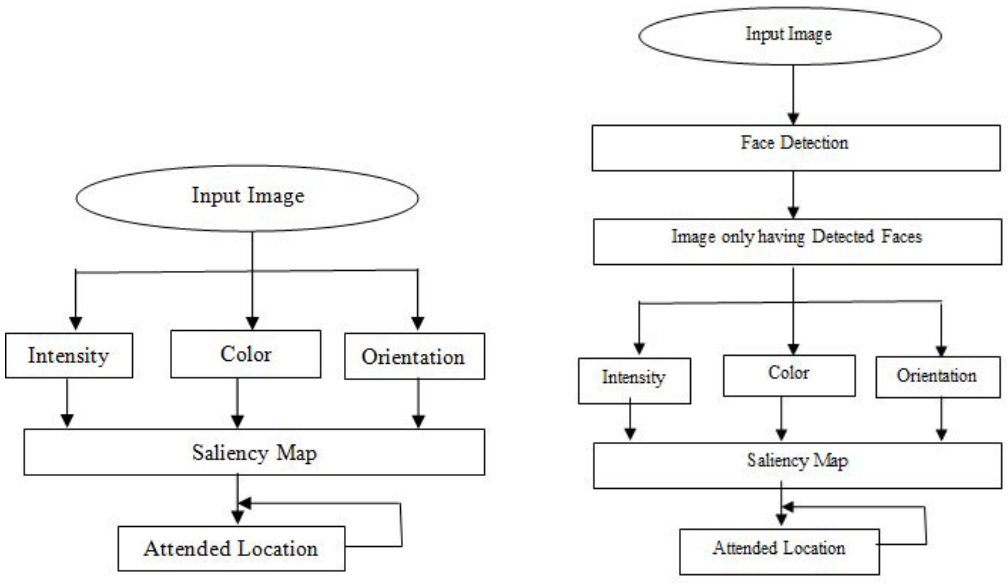

Figure 1. (a) Traditional Saliency Model (b) Proposed Saliency Model

In order to attending faces in a scene (Containing face and non face regions), in most of the traditional work [2,3,4,5,10,14,20] saliency maps have been generated by extracting of low level feature on direct input image. But, in practical, first we just look at all the faces at a glance then after reasoning and interpretation our gaze focuses on the prominent one of the scene. Subsequently

we attend the salient region over this. The proposed saliency framework aim is to map this type of vision modeling.

4. PROPOSED METHODOLOGY

The proposed approach to get saliency values of all the faces in the input image and to determine the most salient face, involves following steps:

a) Face detection

Viola Jones algorithm [21] has been applied to detect the faces. 'Viola Jones Face Detector' is developed by building a 'training classifier' after trained it by features of faces and non-faces as input. Training stage is based on giving a positive and negative set of images. Positive set of images are the faces and negative image set of images have non-faces. During phase of training, features of face and non faces are extracted and stored in a file. In the phase of testing, these features are co-related with the test input image and based on appropriate thresholds it is classified as a face. In the first phase of the algorithm Haar-Like Features [21] are determines. Haar-Like Features are the features of digital image which acquires object characterization. The concept of integral image [22] reduces the computation time of Haar features. Using concept of Integral image [22], the value at pixel (x, y) is calculated as the sum of pixels values of above to the (x, y) and to the left of (x, y). Haar function generates more than 160,000 features but out of many of them are not relevant, the AdaBoost algorithm [14] reduces these by selecting only relevant features. The selected relevant features are Weak Classifier [23]. The linear combination of these weak classifiers build a strong classifier. In the last, cascade classifier [23] is generated which is the composition of several stages, processed with a strong classifier. At each stage, it decides the sub-window inscribe as a face or a non-face.

In Figure 2, Haar features selected by Adaboost algorithm are applied over input face image. The first feature deals with the intensity difference between eyes and its below cheeks area. It shows that the eye region is darker than cheeks area. Second feature is used to determine the intensity difference between the eye region and bridge of the nose.

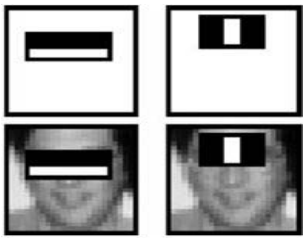

Figure 2. Feature Selection through Adaboost Algorithm [21]

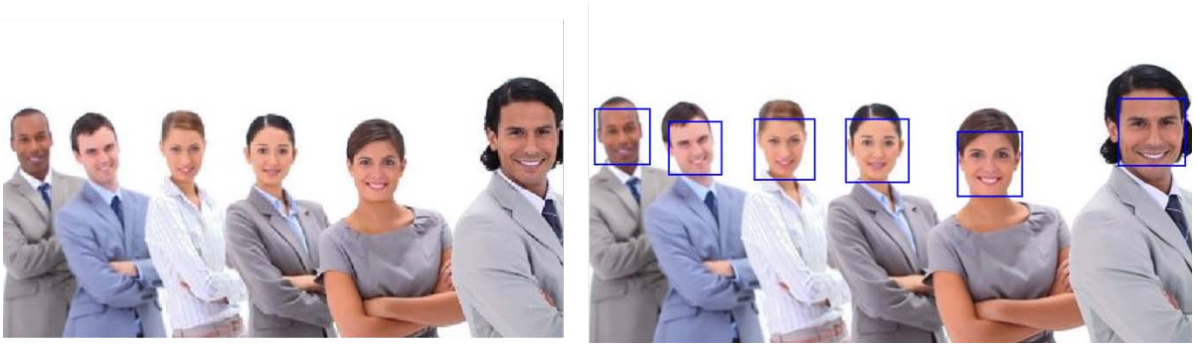

Figure 3. (a) Input Image (b) Detected Faces Using Viola Jones Algorithm[21]

b) Bounding Box Adjustment

Viola jones algorithm [21] detects the faces inscribed in the bonding box. But due to expressions, leaning and tilting face, many times this bounding box does not appears at very relevant places and hence does not cover the full portions of faces Therefore, to cover the missing part of the faces, adjustment in bounding box has been worked out in this work. As per experimented with various face images, the proper adjustment has been made with the size of bounding box. Based on observation and average calculation, for entire coverage of the faces, bounding box (BB_{VJ}) is increased by $1/3^{rd}$ times in +Y direction and $1/4^{th}$ times in –Y, +X, -X directions. After this implementation the tracked face is found to be fully covered inside this new bounding box (BB_{New}). It may be inscribed some extra background regions too but the main goal is to not to missing of the skinny part of the faces. The production of new bounding box(BB_{New}) by enhancing the Viola jone's bounding box(BB_{VJ}) in the all coordinate directions have been done as:

$$BB_{New}(x) = BB_{VJ}(x) + BB_{VJ}(x/4)$$

$$BB_{New}(-x) = BB_{VJ}(-x) + BB_{VJ}(-x/4)$$

$$BB_{New}(y) = BB_{VJ}(y) + BB_{VJ}(y/3)$$

$$BB_{New}(-y) = BB_{VJ}(y) + BB_{VJ}(-y/4)$$

Figure 4. Detected Faces inside New Bonding Box

c) Top Down Attention

As top down attention is target specific and here faces are the target. Therefore, only detected faces are taken into consideration and regions other than faces have been ignored. This has been achieved by selecting all the pixels inside the new bounding boxes (Figure 4) and populating remaining pixels as 0.

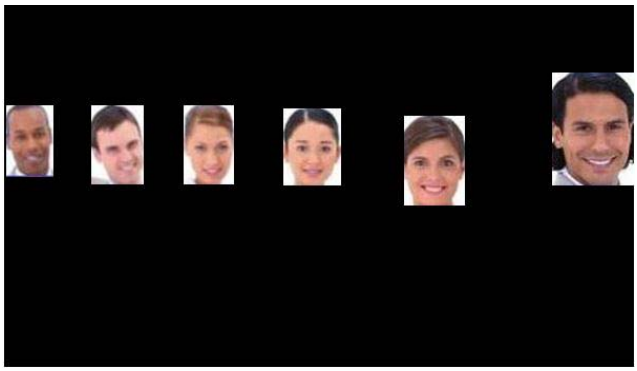

Figure 5. Selected Face Regions After ApplyingTop Down Attention Mechanism

d) Bottom Up Attention

Now, bottom up mechanism has been applied on the output image obtained from top down mechanism. A particular location/object becomes conspicuous when it is dissimilar from its surroundings in terms of one or more low-level features such as, intensity and orientation. In bottom up attention model, various low level features like colors (RGB), orientation, intensity etc. are extracted at various scale. Next, center surrounding difference is calculated and normalization is done to obtain feature and conspicuity maps. Linear combination of these maps results the saliency map.

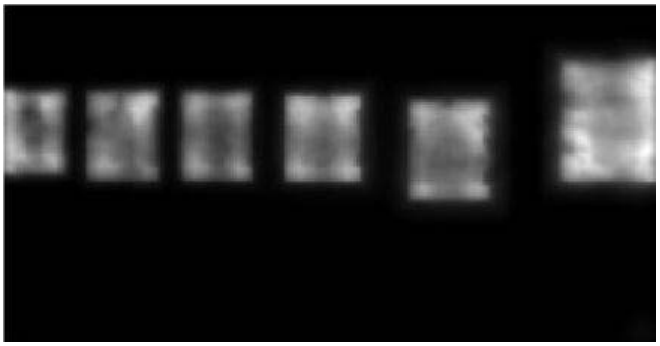

Figure 6. Saliency map exhibiting Face Regions using Bottom Up Attention After Applying [20] over Top Down Attention(Figure 5)

5. EXPERIMENTAL RESULT AND DISCUSSION

For testing this proposal, experiment has been performed over 50 set of colored input crowd images. Each set of input crowd images having 3 to 35 faces have been collected from World Wide Web (WWW). In this work only low level features have been considered, therefore, saliency map have been generated using Context Aware Saliency [20]. Saliency map using ‘bottom up approach’ (general-traditional) and ‘top down followed by bottom up’ (proposed method) have been depicted below.

Figure 7. (a) Input Image (b) Saliency Map using [20] (c) Proposed Saliency Map using[20]

In the above figure 7(b), Saliency map generated by [20], has been shown. The saliency map using is generated for every part of the stimulus (Figure 4(b)). But as our target is to find the salient face in the given images, it means only face information are required. So, the existing approach is not much appropriate for this target specific work. In this saliency map, we can easily observe that many non-face parts also have been depicted as salient. In order to reduce the time and space complexity, our proposed method comes with the top down approach for face detection and the bottom up approach for finding salient faces of the detected faces. In figure 7(c), one can easily observe that the saliency map have only the face information and other irrelevant information of non face regions has been completely ignored by the proposed technique.

6. CONCLUSION

In most of the time, while visiting a scene having human faces, we are giving more importance to human faces rather than other objects. Therefore, there is a need of attending all the available faces in the scene and ignoring the other regions. The traditional approach is not capable to handle such situations, because they treated face and other regions in the same manner. The main goal of this project is to find the salient locations among the faces of the input image. In this work a combined idea of top down and bottom up attention has been proposed and implemented on the input images having multiple humans. Instead of traditional method of estimating saliency map over entire input image, first we have tracked all the available faces of the image and discarded the unwanted portions (non face regions) from the input image. This phase has been accomplished by top down mechanism. In the next phase of experiment, concept of bottom up attention mechanism has been applied to get the saliency map of face area only. This proposed method contributes a little but novel idea to make the computer vision system in the similar way as human vision.

References

[1] R. Pal, R. Srivastava, S.K . Singh and K.K. Shukla, "Computational Models of Visual Attention: A Survey. Recent Advances in Computer Vision and Image Processing: Methodologies and Applications," IGI Global, pp.54-76, 2013

[2] R. K. Kumar, J. Garain, G. Sanyal and D. R. Kisku, "Analysis of Attention Identification and Recognition of Faces through Segmentation and Relative Visual Saliency (SRVS,". In Proceedings of ICIP Procedia Computer Science, pp.756-763, 2015.

[3] C. Koch and S. Ullman,"Shifts in selective visual attention: towards the underlying neural circuitry," Human Neurobiology, Vol. 4, pp.219-227, 1985.

[4] L. Itti and C. Koch, "A Saliency-based search mechanism for overt and covert shifts of visual attention," Vision Research, Vol. 40, pp.1489-1506, 2001.

[5] L. Itti , C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. on Pattern Analysis & Machine Intelligence., pp.1254-1259, 1998.

[6] L. Itti, and C.Koch, "Computational modeling of visual attention". Nature Reviews Neuroscience, vol. 2, no. 3, pp. 194-203, March 2001.

[7] L. Itti, "Models of bottom-up and top-down visual attention." PhD diss., California Institute of Technology, 2000.

[8] N.Ouerhani and H.Hugli, "A model of dynamic visual attention for object tracking in natural image sequences", Computational Methods in Neural Modeling, Lecture Notes in Computer Science, 2686/2003, pp.702–709, 2003b.

[9] A.Bur, A.Tapus, N.Ouerhani, R.Siegwart and H.Hugli, "Robot navigation by panoramic vision and attention guided features", In Proceedings of 18th International Conference on Pattern Recognition, Washington DC, pp. 695–698, 2006.

[10]R.Pal, P.Mitra, and J.Mukhopadhyay, "Icam: Maximizes viewers' attention on intended objects". In Proceedings of Pacific-Rim Conference on Multimedia, pp. 821-824, 2008.

[11]L. Itti, C. Gold, and C. Koch. “Visual attention and target detection in cluttered natural scenes”, Optical Engineering, vol.40, issue 9, pp.1784-1793, 2001

[12]A. M. Treisman and G. Gelade, “A feature-integration theory of attention”, Cognitive psychology, vol.12, issue 1, pp. 97-136, January 1980.