Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Computational-Cognitive model of Visual Attention in Dynamic Environments

M. Shabani, A. Bosaghzadeh*, R. Ebrahimpour

Artificial Intelligence Department, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 04 May 2021
Reviewed 25 June 2021
Revised 05 July 2021
Accepted 07 September 2021

Keywords:
Visual Attention
Dynamic Visual Attention
Bottom-Up Attention
Visual Saliency
Human Eye Fixation

*Corresponding Author's Email Address:
a.bosaghzadeh@sru.ac.ir

Abstract

Background and Objectives: Visual attention is a high-order cognitive process of the human brain that defines where a human observer attends. Dynamic computational visual attention models are modeled on the behavior of the human brain. They can predict what areas a human will pay attention to when viewing a scene such as a video. However, several types of computational models have been proposed to provide a better understanding of saliency maps in static and dynamic environments; most of these models are used for specific scenes. In this paper, we propose a model that can generate saliency maps in various dynamic environments with complex scenes.

Methods: We used a deep learner as a mediating network to combine basic saliency maps with appropriate weighting. Each of these basic saliency maps covers an essential feature of human visual attention, and ultimately the final saliency map is very similar to human visual behavior.

Results: The proposed model is run on two datasets, and the generated saliency maps are evaluated by different criteria such as ROC, CC, NSS, SIM, and KLdiv. The results show that the proposed model has a good performance compared to other similar models.

Conclusion: The proposed model consists of three main parts, including basic saliency maps, gating network, and combinator. This model was implemented on the ETMD dataset, and the resulting saliency maps (visual attention areas) were compared with some other models in this field by evaluation criteria, and their results were evaluated. The results obtained from the proposed model are acceptable, and based on the accepted evaluation criteria in this area; it performs better than similar models.

Introduction

In recent years, many scientific studies have been conducted with the aim of understanding and modeling effective mechanisms in predicting the areas of human visual attention. Neurobiologists have also conducted extensive studies and research to show how neurons adapt to display objects [1] better. Inspired by these studies, researchers and scientists in the field of robotics and computer vision have tried to create computational models that mimic the human visual attention, more attractive areas of the human visual system to predict [2], [3].

The main challenge of computational modeling of visual attention is how and why to predict areas of the image or video as important areas from the perspective of human visual attention. In this regard, several factors have been introduced to influence and increase the attractiveness of some areas, the most obvious of which

is the inspiration of the anatomy and function of the primary human visual system [2], [3], [24].

Visual attention includes all the factors that affect the selection mechanisms of important areas that can be influenced by environmental motivational factors such as the bottom-up approach or the search to find the desired areas (or top-down approach) [4], [30]. A quarter of the brain is involved in the processing of the human visual system (HVS) information; The human visual system receives a large amount of visual data from the environment, Then ultra-fast and reliable focused on areas with more attractive areas [6].

Computational modeling of this intelligent behavior remains a challenge [7]; in addition, processing this data without the help of intelligent mechanisms is difficult to reduce the amount of incorrect visual data. High-level, complex cognitive processes, such as object recognition or scene interpretation, can be transformed to select important areas of the image or video. These mechanisms of conversion and recognition of more important areas are known as visual attention [8].

In the human visual system, attention is facilitated by the retina, which includes the central Fovea with high resolution and the surrounding area with low resolution. One of the effective factors in the process of visual attention is the center-surround difference which is inspired by neural responses in the lateral geniculate nucleus (LGN), and the visual cortex (V1), another effective factor in visual attention is the center-biased elements; they attract more visual attention [9]; also the field of view, which is the response field of neurons, is effective in human visual attention [28]. In fact, visual attention provides a structure by which to focus more on the important parts of the scene that contain more important information and to gather more accurate information; these important parts of the scene are called visual saliency. The term "saliency" is often used in bottom-up computations. In the human visual system, saliency is encoded by superior colliculus neurons located in the front of the thalamus [4]. So far, various models have been proposed to detect and predict visual saliency, which, regardless of the approach used in the model, try to more accurately predict salient points to be closer to attractive areas in the human visual system [26].

Visual attention models are directly or indirectly inspired by cognitive concepts. There are several criteria for classifying and distinguishing models of visual attention and better understanding it, which determine the scope of application of different models (see [4] for more study). Here, in addition to the bottom-up approach, we focus on models of visual attention that can compute saliency maps from any image (or frame) input; although some models evaluate static images, there may be some important points of interest when the object is moving in the video. Hence, visual attention models can be divided into two categories: visual attention models in static environments and visual attention models in dynamic environments.

A. Models Based on Visual Attention in Static Environments

The basic model of Itti et al. [2] uses three feature channels as color, intensity, and orientation. First, an input image is inserted into a Gaussian pyramid, and each level of the pyramid is broken down by sigma into color channels (including red, green, blue, yellow), intensity, and local orientations. From these channels, "feature maps" are created, aggregated, and re-normalized, creating "conspicuity maps" for each channel. The conspicuity maps are then combined linearly to form a saliency map. Finally, based on the Winner-Take-All network model, the more saliency points are selected and then Inhibition and transferred to another salient point so that all the salient points are selected. Finally, according to the Winner-Take-All network model, the more salient points are selected and then inhibited and transferred to another salient point so that all the salient points are selected. This model better displays areas of the image that have center-surround differences but does not cover features such as center-biased and global contrast [27].

Zhang et al. [10] propose a visual attention model called SUN, based on saliency using natural statistics. The information theory model is based on the assumption that local saliency computations are used to maximize sampling information from its surroundings. Bruce and Tsotsos [11] showed that Shannon self-information is used to compute the importance of image regions in a neural circuit, indicating a close relationship with the circuit in the primary visual cortex. In fact, the work of this model is based on closely related values through self-information. They leave out another information scene by selecting sections that contain more useful information. Although this model focuses on uncommon parts of the image, it does not cover features such as center-biased. Harel et al. [12] introduced the graph-based visual saliency model. In this model, activation maps are first created in specific channels of the feature, normalized for greater conspicuity, and then combined. The algorithmic nature of this model is "center-biased," and when the algorithm is implemented, the nodes tend to the central node. This model is simple and biologically acceptable and operates in parallel. The nature of the algorithm covers the tendency to center well but does not examine the salient features globally throughout the image. Hou and Zhang [8] offer a model that is independent of features, categories, or other forms of prior knowledge of objects

and use the power of the Fourier spectrum to detect saliency. By analyzing the logarithm of the input spectrum of an input image, the spectral residue of an image is extracted in the spectral domain. This model proposes a rapid method for constructing a saliency map in a spatial domain and examines the saliency globally but does not cover features such as center-surround difference and center-biased.

To better adapt to the visual attention mechanism in the human brain, temporal and motion information is also needed, which the mentioned models cannot extract. Hence, the models that use this information are called models based on visual attention in dynamic environments; In the following, we will explain these models.

B. Models Based on Visual Attention in Dynamic Environments

Since most of the scenes in front of us are dynamic and the movement of objects in the scene can have a significant effect on attracting the audience's visual attention and the features used in static scenes do not cover this issue, recently, models have been proposed that have also used the effect of motion feature in predicting visual attention. Rogalska and Napieralski [15] proposed a model that uses four features of global contrast, distance from the center of the frame, the human face, and movement in natural scenes with or without the human face to extract visual attention. This model has based on the constant weighting of features that improper weighting has a direct impact on the output. Hongfa et al. [7] proposed a deep integration-based model consisting of three stages of deep feature extraction, deep feature integration, and saliency prediction to detect the salient object in the videos; The inputs of this model are video frames and optical image. Then, deep multi-level features are hierarchically integrated using an integrated network, and finally, with the entry of objects boundary within the frame, the final saliency map is produced. Despite the hierarchical integration in this model, the effect of feature boundaries is very effective and affects visual attention points. Wang et al. [16], [17] proposed a framework that segmented video by visual attention and extracted objects from it. The input frame is converted into three images, including the segmentation map on the superpixels, the Spatio-temporal edge map obtained from the static edge probability, and the gradian size of the optical flow map. For each superpixel, the probability of objects is obtained, and the saliency estimate within each frame and between two consecutive frames is determined, and by combining them, the final saliency map is produced. In this model, for frames where the superpixels are not well defined, the selection of areas of attention is associated with lower performance. Meijun et al. [18] proposed a model that predicts visual attention in films through a convolutional neural network and an optical flow-based method; first, a deep learning framework is used to extract the spatial properties of the frames, then the temporal properties of moving objects in the video frames are computed through the optical flow. By merging these two sets of features, a combined spatial-temporal feature set is obtained as input to a support vector machine (SVM) to predict visual attention. In this model, the effect of feature extraction by the network has a great impact on the output of the model and affects the proper selection of saliency areas. Koutras and Maragos [19] propose a multifunctional Spatio-temporal network called SUSiNet that can jointly solve Spatio-temporal problems of saliency estimation, performance recognition, and video summaries. The proposed network uses an integrated architecture that includes a general and specific task layer and uses the same video input to generate different types of outputs, i.e., saliency maps or classification labels. The effect of observer tags is very important in this model and even affects the motion feature.

Due to the fact that important information can be in several consecutive frames and also in some scenes the movement of stimuli along with other features has a significant effect on attracting the audience's visual attention, the motion feature is important; Therefore, in this research, with a bottom-up approach and based on visual characteristics in static and dynamic environments, visual saliency is extracted from video frames, and based on that, the position of important areas is computed, and the areas of visual attention are predicted [29]. Also, for better adaptation to the human visual system, two features of skin and face are used. The model is also designed to produce an efficient final saliency map with deep network training for automatic and appropriate weighting.

Unlike previous models, the proposed model uses several different features. Also, the proposed model uses a gating network, which provides different weights (between 0 and 1) based on the type of frame, while previous models use fixed weights for composition. In addition, using a combination of top-down and bottom-up models in the proposed model compared to other existing models that usually use one of the models is another advantage of the proposed model.

Technical Work Preparation

Our proposed model consists of three main parts, including basic saliency maps, gating network, and combiner. Basic saliency maps are based on four basic models and two basic features to produce an efficient visual attention model in a variety of video inputs. Fig. 1 shows an overview of the proposed model; as shown in

Figure, each input frame contains saliency maps to fit the base models. After each frame enters the gating network, weighted base saliency maps are assigned and combining according to the weight (between 0 and 1), the final output includes a weighted combined saliency map.

Also, before performing the combination, normalization operations are performed on the saliency maps. The saliency maps generated for each frame are sequentially formed to form a video that, as computational visual attention in a dynamic environment, estimates areas that are important to the human visual system in the video.

According to Fig. 1, the proposed model has three main sections, each of which is described below:

A. Basic Models and Features Used

Since visual attention models have unique features in selecting salient areas and one model does not have the best performance in all scenes, the proposed model tries to use the advantages and strengths of the models used to select salient areas. One of the determining factors for estimating salient areas in visual attention models is the features used and how they are combined. In our proposed model, four basic models and two features are used as basic saliency maps so that they can efficient for a variety of complex scenes, with high or low change speeds, containing images of human faces and etc. In the following, the basic saliency maps and the reason for their use in the proposed model are explained.

1. Hou and Zhang spectral residual model [8] :

This model is independent of features, categories, or other forms of prior knowledge of objects; instead of processing frames in the spatial domain, it computes saliency in the frequency domain and quickly and powerfully examines saliency globally.

Using this model helps us to identify salient areas in the whole image globally. An example of the output of this model can be seen in Fig. 2-b.

2. Computational visual attention model of Wang et al. [16], [17]:

For each input frame, this model uses a combination of the temporal-spatial saliency map, segmentation map based on superpixels, and motion information in several consecutive frames and produces a saliency map based on dynamic environments.

In this model, the final saliency map is very much influenced by the movement of saliency areas, so we have used this model to extract motion information in two consecutive frames. An example of the output of this model can be seen in Fig. 2-c.

3. Graph-based model of Harel et al. [12]:

This model uses the features of color, intensity, and orientation and extracts the salient areas based on graph and Markov. It then combines the saliency maps and produces the final saliency map. This model is simple and biologically acceptable. This model powerfully predicts the human gaze, and the nature of its algorithm is such that the saliency map produced tends to be centered.

We use this model as a basic model because of its biological acceptance and its tendency to be central, which also applies to the human visual system. An example of the output of this model can be seen in Fig. 2-d.

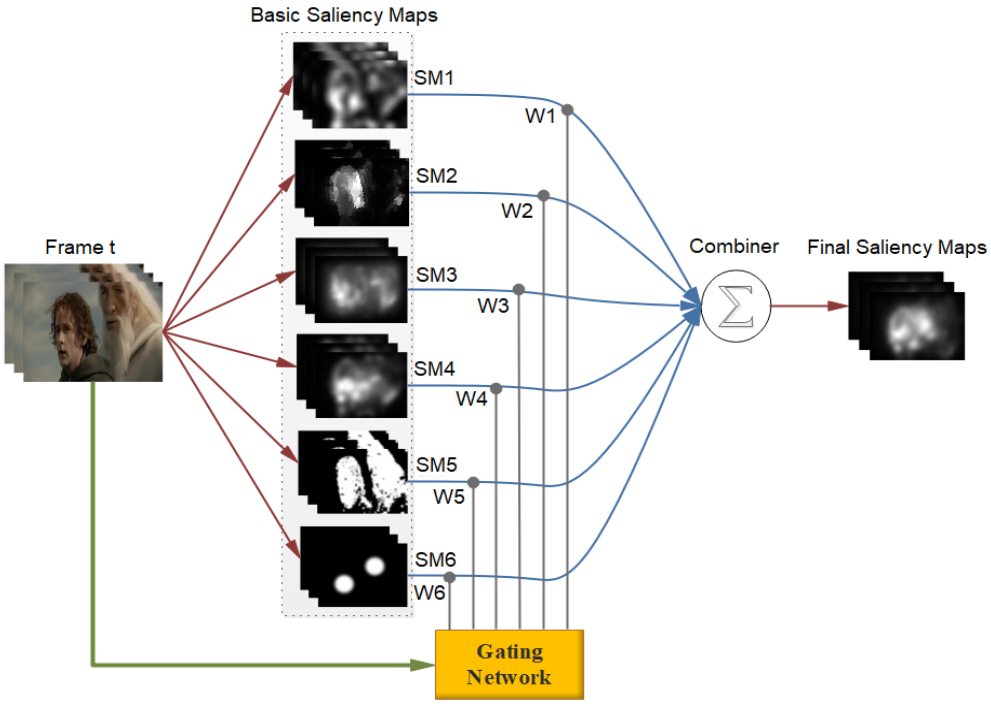


Fig. 1: Block diagram of the proposed model, including Basic Saliency Map, Gating Network, and Combiner.

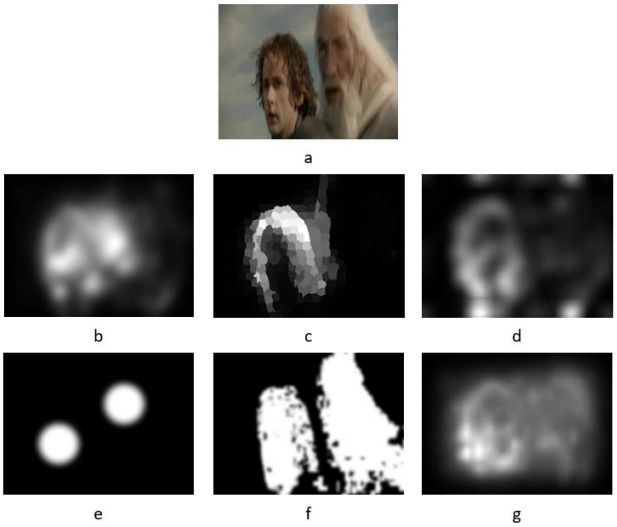


Fig. 2: Examples of basic saliency maps for a frame, a: main frame, b: Harel et al. [12], c: Wang et al. [16], [17], d: Hou and Zhang[8], e: recognizing areas with faces [21], f: skin diagnosis[20], g: Itti et al.[2].

4. Itti et al. Model [2]:

Itti et al. proposed a model that uses three feature channels: color, intensity, and orientation. This model is the basis for later models and a standard benchmark for comparison.

The model of Itti et al. is inspired by cognitive concepts, is very much in line with the mechanism of the human visual system, and shows well the intensity, color, and center-surround difference; therefore, we used this model as a base saliency map in the proposed model. An example of the output of this model can be seen in Fig. 2-c.

5. The basic feature of skin diagnosis:

Many of the scenes have areas of it including human skin, and because experimentally, human visual attention is biased towards skin color, in the proposed model, we used the saliency map obtained from this feature along with other saliency maps; also, for further adaptation with the human visual system, we use this feature that is kind of top-down feature. An example of the output of this model can be seen in Fig. 2-f. How to compute skin areas is based on color differences.

By first obtaining the difference between red and green, blue and yellow; Calculates the H and S values in the HSV space based on the difference obtained, then checks the H and S values for all frame pixels if they are in the desired range for the skin (H = 3.14> H> 1.930 <S <130) Sets the pixel value to 1 (or the pixel value of the original image) and otherwise to 0 [20].

6. Basic feature of recognizing areas with faces:

Like the skin, the human face is repeated in many scenes because the human visual system is biased towards the face; the marked areas based on the face have been used as another basic saliency map in the proposed model. Like the skin, in addition to adapting more to human vision, we also use the top-down attention feature. An example of the output of this model can be seen in Fig. 2-g, which show how to compute the saliency map from areas with human faces based on the trained classifier Haar Cascades [21], after detecting the location of the face in the method expressed in [21], we draw a circle from the facial center to the radius proportional to the detected face. Then, the facial area takes 1, and the rest are 0, and applies a Gaussian filter (as Gaussian smoothing) to smoothen the edge of the face.

To better understand the basic saliency maps, Fig. 2 shows an example of a frame containing these saliency maps used in the proposed model.

B. Learning Process and Gating Network

After determining the basic saliency maps, it is important to select and participate in the basic saliency maps to produce the final saliency map. To do this a deep learning model called Alexnet was used to be able to learn in different scenes with proper weighting to the basic saliency maps, Provide the basis for the production of an efficient saliency map. The process of learning and creating a gating network can be seen in Fig. 3.

The important issue in this section is to set the parameters and how to train this network. For training, we set the number of classes equal to the number of basic saliency maps and based the class selection on the highest value of CC. Thus, for the frames to be used as training data, we first obtain the basic saliency maps suitable for each model and then compute the CC value of the frames for each saliency map. Then, for each frame, the maximum value of CC determines the basic saliency map. For example, for a frame (t), we will have six types of basic saliency maps, so each of the basic saliency maps with the highest CC value is considered as the desired saliency map for a frame (t). Finally, we will have a class number (or the base protrusion map number) for each frame (t).

After determining the class number of the saliency map for each frame, the deep network is trained based on the frames and their class number. Finally, the network trained as a gating network weighs the basic saliency maps for each input frame to combine and produce the output saliency map. The weight (between 0 and 1) assigned to each basic saliency map is computed according to the Softmax function (in layer 24 of the Alexnet deep network) and obtained from (1). the above steps can be seen in Fig. 3.

$$\sigma(Z)_j = \frac{e^{Z_j}}{\sum_{k=1}^{K} e^{Z_k}}, for\ j = 1, .., K \quad (1)$$

In (1), a K-dimensional vector of real numbers such as Z is received as input, and a K-dimensional vector of σ (Z) of real values [0,1] is used as an output, the sum of its components is 1.

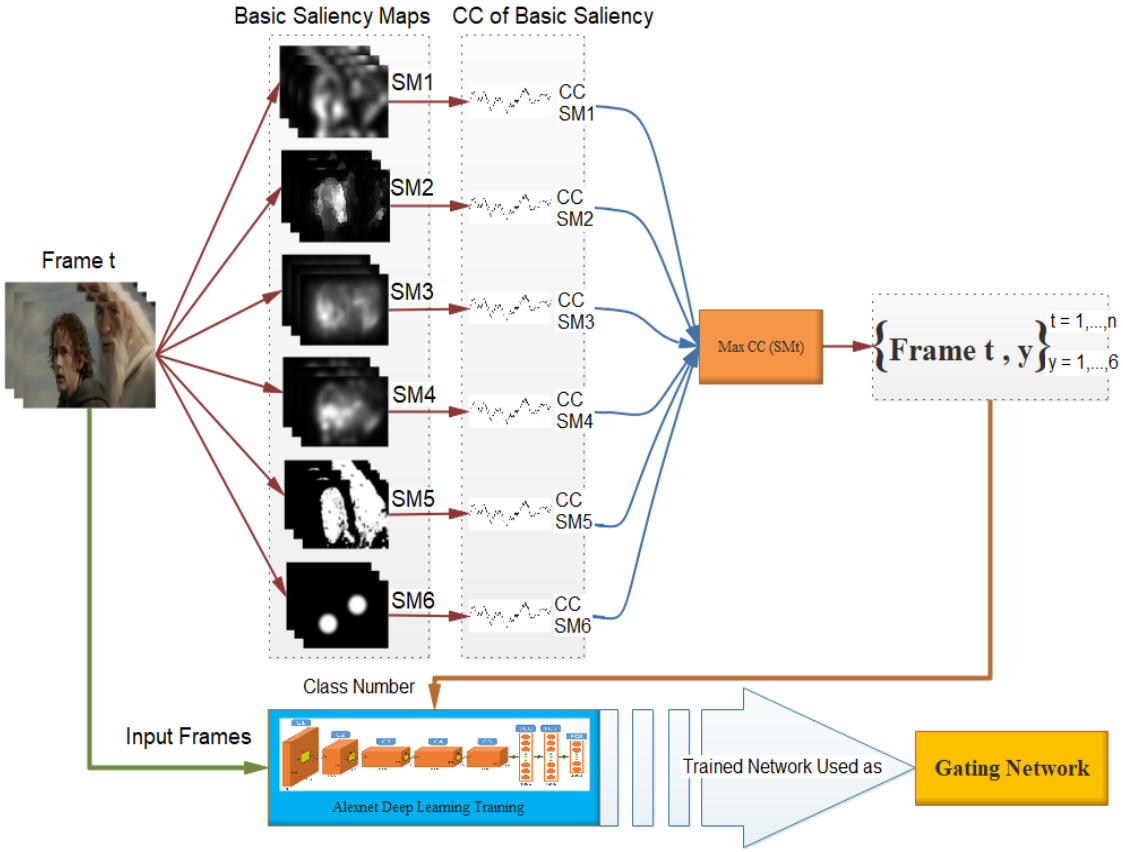


Fig. 3: Deep network training block diagram for use as a gating network.

C. Combiner

The final step of the proposed model is the combination of basic saliency maps. In this step, for each input frame (t), the gating network assigns weight (between 0 and 1) to each basic saliency map, then according to (2), based on the given weight, the saliency maps are combined linearly, and the final saliency map is created.

$$SM_{Output} = \sum_{i=1}^{6}(W_i \times SM_i) \quad (2)$$

where SM_{Output} is the final saliency map, i is the frame number, and W_i is the weight given by the gating network to any saliency map SM_i. Prior to compounding, the saliency maps are normalized using the method used in[2]. Finally, after computing the final saliency map for each frame, it is put together, and the video creates computational-cognitive video attention based on dynamic environments.

Results and Discussion

For evaluation, the proposed model with similar models uses the standard ETMD dataset, with diverse, complex, and dynamic scenes. For this work, the final saliency maps corresponding to each clip of the dataset are examined based on standard evaluation criteria.

To train the network, each video is divided into two parts; the first part is used to train the gating network, and the second part is used to test and evaluate the model. Because we have trained and evaluated the network separately on each video, the accuracy of the network is different; an average is equal to 94%. In fact, the proposed model predicts saliency maps for frames that were not observed in the training process.

A. Dataset

There are several eye movement datasets of still images (for studying static visual attention) and videos (for studying dynamic visual attention). In [4], some existing datasets are listed. In this study, the ETMD dataset is used [22].

This dataset consists of areas tracked based on human visual attention, which includes video clips from Hollywood movies. This dataset contains relatively long videos containing complex scenes. In this dataset, visual attention is designed based on both low-level features (such as intensity, color, and temporal-spatial energy) and high-level (face recognition [25], etc.). This dataset can be useful in many visual attention tests because, according to the evaluation criteria, it has achieved good performance in the practical prediction of the human visual system [22].

Fig. 4 shows an example of the fixation of the human eye in a frame (for the color and gray versions of

movies). As you can see, there is a global correlation between different users and the color or gray version in each video frame. This correlation expresses the viewers' consensus in focusing their visual attention on the same areas. In most cases, the fixation points of all viewers are generally close to each other.

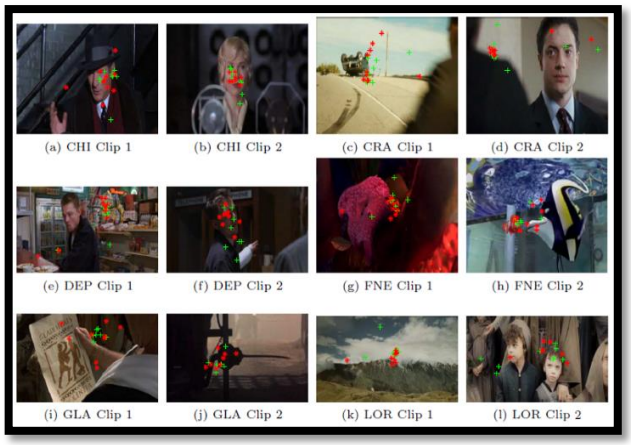


Fig. 4: Examples of human eye fixation points in frame number 500 for every 12 movie clips. Green + dots are compared to the color version of each video, and red * Dots are related to the gray version of the video [22].

B. Experiments and evaluation results

To perform the experiments, we first run the desired models and the proposed model on the video in the mentioned dataset and obtain the saliency maps appropriate to each model. The saliency maps obtained from the proposed model are compared and evaluated with other models. The basis of comparison is evaluation criteria, such as CC, NSS, KLdiv, ROC, SIM [4], [23]. The final value of each criterion is equal to the average of that criterion in the total frames used in each experiment.

To run models on videos, we first need to convert them to frames. It should be noted that the proposed model is trained on half of the frames, and then the saliency maps produced from the other half are used as test and evaluation frames and comparisons with other models. An example of the output of different models with GSM and the main frames for a frame can be seen in Fig. 5.

As shown in Fig. 5, for each of the selected videos from the dataset, the saliency map is computed and displayed according to each model.

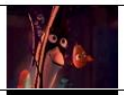
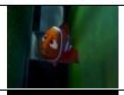
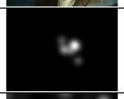
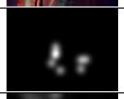
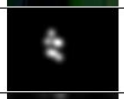
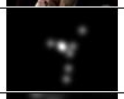
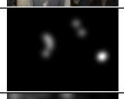
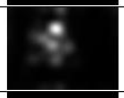
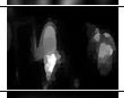
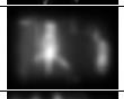
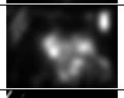
Column	1	2	3	4	5	6
Clip Name	CRA Clip 2	FNE Clip 1	FNE Clip 2	GLA Clip 1	LOR Clip 1	LOR Clip 2
Frame Number	02441	01498	01870	02104	02367	00417
Frames						
GSM						
Hou&Zhang[8]						
Wang[17]						
GBVS[12]						
Itti&Cokh[2]						
Skin[21]						
Face[20]						
Proposed Model						

Fig. 5: Samples of the dataset frame and along with the saliency map produced along with the human fixation points.

In Fig. 6, due to the lack of a human face image, there is no saliency map equivalent to the face and skin for the desired frame; however, the proposed model has an optimal saliency map. The reason for this is proper weighting by the gating network. In fact, proper weighting by the gating network reduces the effect of the prediction error of the base models.

In the dataset used, the video animation in some frames has a face or skin saliency map, which can be seen in Fig. 5, columns 2 and 3. The reason for this is the type of animation design that has features of the face or skin (such as the geometric shape of the limbs, skin, etc.), so for such frames, there are saliency maps based on the face and skin. In addition, in Fig. 5, the saliency of the last column is very similar to the face saliency; because of the weight (between 0 and 1) of the Gating Network, the saliency map of the face has taken more weight; it is also adapted to GSM.

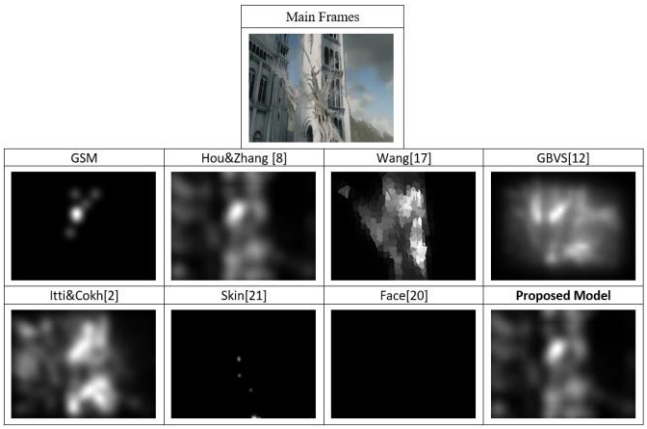


Fig. 6: An example of a dataset frame with human fixation points, without a saliency map for some models.

For the experiment, six videos from the ETMD dataset were used to prepare the saliency maps by the proposed model. That's why these six videos have been selected, which include frames with different characteristics, including the human face, skin, text, complex scenes, high and low speed in scene changes, and more. These reasons make the performed experiments, and the output of the proposed model more in line with the real world, and the results of the experiments are more reliable.

In Table 1 to Table 6, the results obtained based on the mentioned evaluation criteria can be seen. It should be noted that in the experiments, the selected frames for each video were from the test section and were not used in the training process of the proposed model.

As can be seen in Table 1 to Table 6, in general, the results obtained in each of the videos are better for the proposed model than similar models. In some cases, the proposed model is in the second order with a slight difference (first order, bold and second-order, gray color). The reason for this includes the characteristics of the evaluation criteria and the sensitivity in the distribution of salient areas within the saliency map, which causes the model to be second in some criteria. For example, in Table 1, the GLA Clip 1 video is second only to the proposed model in terms of similarity and divergence KLdiv.

Table 1: Values obtained from the visual attention according to the evaluation criteria selected frames of video datasets ETMD, GLA Clip 1

Method	Metrics				
	CC	NSS	KLdiv	SIM	ROC
Hou&Zhang [8]	0.0321	0.1607	3.2693	0.1065	0.5118
Wang [17]	0.1281	0.4766	3.9677	0.1499	0.5268
GBVS [12]	0.1998	0.8291	**2.3260**	**0.1617**	0.5722
Itti&Cokh [2]	0.1468	0.6078	2.5643	0.1388	0.5688
Skin [21]	0.0361	0.1071	4.4577	0.1050	0.5249
Face [20]	0.0133	0.0484	25.4884	0.0138	0.5061
Proposed Model	**0.2036**	**0.8321**	2.3486	0.1585	**0.5752**

Table 2: Values obtained from the visual attention according to the evaluation criteria selected frames of video datasets ETMD, CRA Clip 2

Method	Metrics				
	CC	NSS	KLdiv	SIM	ROC
Hou&Zhang [8]	0.1145	0.3990	3.0529	0.1422	0.5247
Wang [17]	0.2241	0.7808	3.1803	0.2106	0.5662
GBVS [12]	0.2506	0.9447	2.1690	0.1870	0.5863
Itti&Cokh [2]	0.2236	0.8237	2.3019	0.1724	**0.5994**
Skin [21]	0.1069	0.3533	18.6394	0.1190	0.5384
Face [20]	0.0183	0.0665	25.2002	0.0171	0.5089
Proposed Model	**0.2790**	**0.9688**	**2.1532**	**0.2126**	0.5877

Table 3: Values obtained from the visual attention according to the evaluation criteria selected frames of video datasets ETMD, FNE Clip 1

Method	Metrics				
	CC	NSS	KLdiv	SIM	ROC
Hou&Zhang [8]	0.3331	1.0488	2.2481	0.2368	0.5558
Wang [17]	0.4005	1.2348	2.4355	**0.3177**	0.6001
GBVS [12]	0.4010	1.3857	1.8612	0.2250	0.6254
Itti&Cokh [2]	0.3157	1.1236	2.0267	0.2146	0.5979
Skin [21]	0.1900	0.4857	16.2675	0.0017	0.5125
Face [20]	0.0528	0.1333	24.5557	0.1491	0.5134
Proposed Model	**0.4178**	**1.4206**	**1.8080**	0.2431	**0.6262**

Table 4: Values obtained from the visual attention according to the evaluation criteria selected frames of video datasets ETMD, FNE Clip 2

Method	Metrics				
	CC	NSS	KLdiv	SIM	ROC
Hou&Zhang [8]	0.1904	0.7010	2.4800	0.1819	0.5392
Wang [17]	0.2726	0.9470	2.8603	**0.2501**	0.5861
GBVS [12]	0.3333	1.2186	1.8546	0.2293	0.6279
Itti&Cokh [2]	0.2873	1.0962	2.0089	0.2089	0.6271
Skin [21]	0.0603	0.0025	15.8132	0.0009	0.5151
Face [20]	0.0026	0.0323	25.6604	0.0091	0.5035
Proposed Model	**0.3382**	**1.2303**	**1.8485**	0.2338	**0.6294**

Table 5: Values obtained from the visual attention according to the evaluation criteria selected frames of video datasets ETMD, LOR Clip 1

Method	Metrics				
	CC	NSS	KLdiv	SIM	ROC
Hou&Zhang [8]	0.1921	0.6884	2.5111	0.1499	0.5449
Wang [17]	0.3396	1.1661	2.2526	**0.2489**	0.6024
GBVS [12]	0.3692	**1.3686**	1.9935	0.2059	0.6326
Itti&Cokh [2]	0.2395	0.9339	2.3517	0.1622	0.5979
Skin [21]	0.2822	0.9091	9.7223	0.2202	0.6217
Face [20]	0.0981	0.3672	21.8202	0.0799	0.5480
Proposed Model	**0.3767**	1.3403	**1.9631**	0.2245	**0.6582**

Table 6: Values obtained from the visual attention according to the evaluation criteria selected frames of video datasets ETMD, LOR Clip 2

Method	Metrics				
	CC	NSS	KLdiv	SIM	ROC
Hou&Zhang [8]	0.1399	0.5043	2.4463	0.1602	0.5281
Wang [17]	0.3150	1.0731	2.4419	**0.2694**	0.6058
GBVS [12]	0.3543	1.2617	1.8319	0.2311	0.6382
Itti&Cokh [2]	0.2639	0.9650	2.0859	0.1911	0.6102
Skin [21]	0.2494	0.8406	6.3158	0.2285	0.6423
Face [20]	0.1255	0.4131	21.6260	0.1063	0.5445
Proposed Model	**0.4022**	**1.3717**	**1.7223**	0.2559	**0.6644**

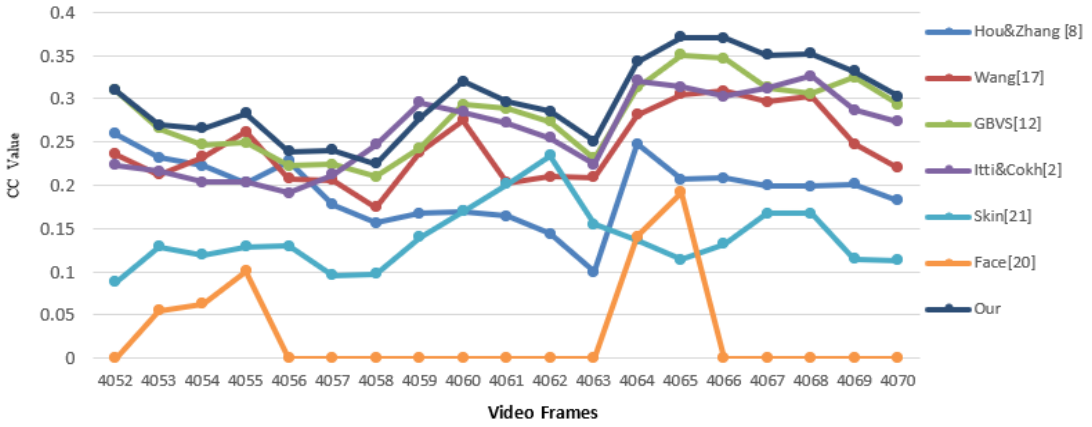


Fig. 7: Chart of values of visual attention evaluation criteria for selected frames of LOR Clip 1 video in ETMD dataset.

In order to better display and understand the changes in visual acuity, during the framing process, multi-frame CC values were plotted for LOR Clip 1 and FNE Clip 2 video from the ETMD dataset, as shown in Fig. 7 and Fig. 9. Respectively in consecutive frames, the performance of the models is different, and one model alone does not have the best result in all frames, so with the optimal combination made in the proposed model, the produced saliency map (Our) has the best performance.

Also, the values of the gating network to each of the saliency maps for produce the final saliency map in the two videos can be seen in Fig. 8 and Fig. 10.

As can be seen, for each frame, the proposed model in Fig. 7 and Fig. 9 is generated respectively from the set of weights assigned by the gating network in Fig. 8 and Fig. 10. In the set of frames, the highest weight is not assigned to only one specific base model, and the score assigned to the models varies according to each frame. This indicates the use of the positive features of each base model by assigning the highest weight, which is weighted by the gating network.

According to Fig. 7 to Fig. 10, models with a larger CC value have received higher values through the gating network and, in fact, have a greater share in the production of the final saliency map.

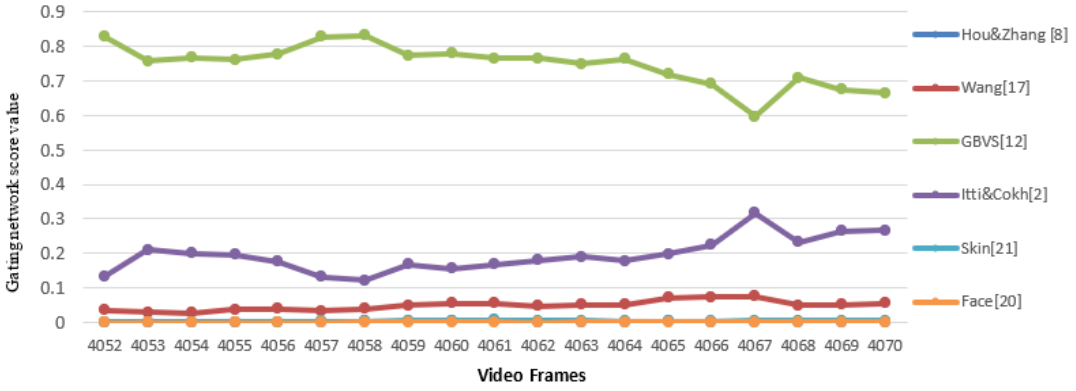


Fig. 8: Graph the values of the gating network score for the selected frames of the LOR Clip 1 video in the ETMD dataset.

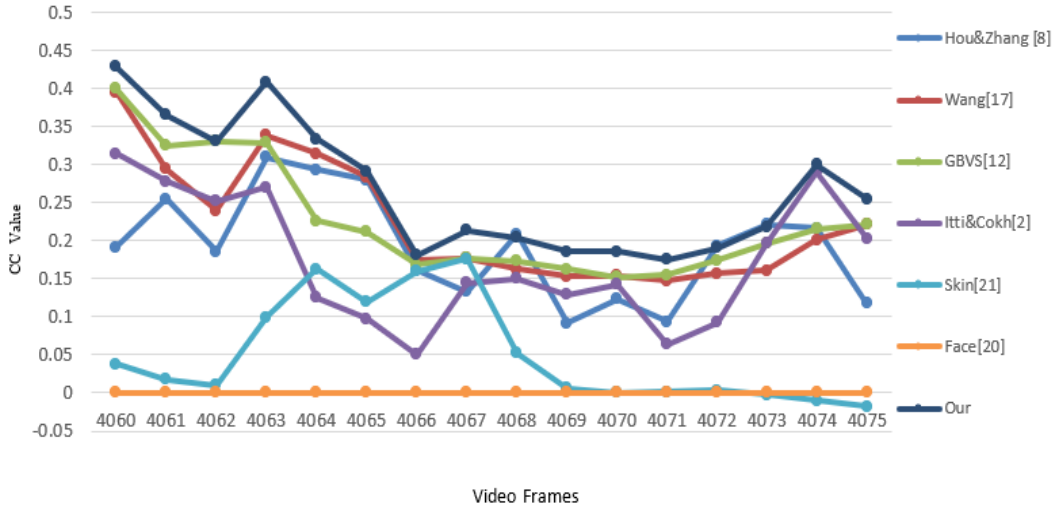


Fig. 9: Chart of values of visual attention evaluation criteria for selected frames of FNE Clip 2 video in ETMD dataset.

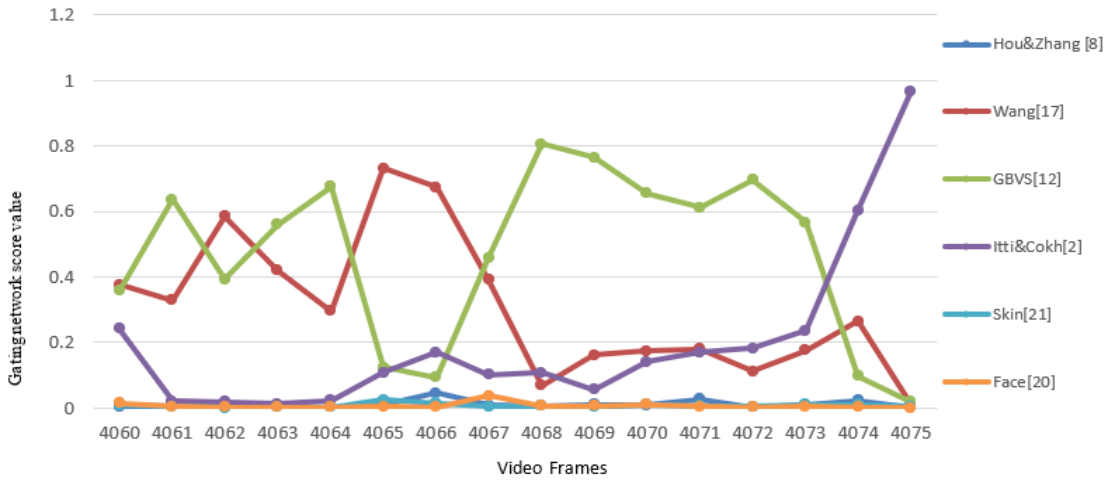


Fig. 10: Chart of values of visual attention evaluation criteria for selected frames of FNE Clip 2 video in ETMD dataset.

Visual attention covers a wide range of applications in a variety of fields; Advances in this area can greatly help solve difficult problems of visual challenges such as interpreting complex and irregular scenes and recognizing the subject.

As can be seen in Table 1 to Table 6, similar models do not have the same performance, and it is not possible to select a model as the best model in all conditions, and the models give different outputs according to different conditions, but the proposed model has good performance and offers acceptable results in various conditions due to intelligent weighting.

On the other hand, several factors affect the bottom-up visual attention that has been discovered by researchers in this field; we also used effective features to predict areas with higher levels of attention from the perspective of the human visual system; However, there are other factors that need to be considered to help reduce the difference between human observations (GSM) and increase the predictive accuracy of computational models.

Also, due to the existence of different models in the field of visual saliency detection, it is still necessary to provide a model with a better and deeper study that covers the weaknesses of current models and provides a comprehensive solution. Our main approach is also comprehensive in the proposed model and fits different scenes with various features, and as can be seen in the results and evaluations, it has a good performance compared to similar models. However, the issue of computational visual attention can be explored in future work by those interested in this field.

Conclusion

In this study, a bottom-up visual attention model based on basic saliency maps was proposed. The proposed model consists of three main parts, including basic saliency maps, gating network, and combinator.

This model was implemented on the ETMD dataset, and the resulting saliency maps (visual attention areas) were compared with some other models in this field by evaluation criteria, and their results were evaluated. As can be seen in the results section, the results obtained from the proposed model are acceptable, and based on the evaluation criteria accepted in this area, it performs better than similar models.

Most bottom-up research is based on visual attention, while task-based visual attention requires further research and development of precise computational principles. For future research, the development of models that can bring computational time closer to real-time can be considered, especially in interactive, complex, and dynamic environments. In addition, there is no principled computational understanding of cover vision, which should be clarified in the future. Also, evaluating and understanding the concepts of images as well as more precise computational principles in this field can be considered in the future.

Author Contributions

M.shabani performed the experiments and interpreted the results. A.bosaghzadeh and R. Ebrahimpour worked on conceptualization and developed the methodology. All authors discussed the results and contributed to the writing of the manuscript.

Acknowledgment

This study has been supported by Shahid Rajaee Teacher Training University.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, ethical issues, including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy, have been completely witnessed by the authors.

Abbreviations

SM	Saliency Map
GSM	Ground Truth Saliency Map
HVS	Human Visual System
LGN	Lateral Geniculate Nucleus
SVM	Support Vector Machine
ETMD	Eye Tracking Movie Database
CC	Correlation Coefficients
NSS	Normalized Scanpath Saliency
SIM	Similarity
KLdiv	Kullback-Leibler divergence
ROC	Receiver Operating Characteristic

References

[1] S. Treue, “Neural correlates of attention in primate visual cortex,” Trends Neurosci., 24(5): 295-300, 2001.

[2] L. Itti, C. Koch, E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., 20(11): 1254-1259, 1998.

[3] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, “Modeling visual attention via selective tuning,” Artif. Intell., 78: 507-545, 1995.

[4] A. Borji, L. Itti, "State-of-the-art in visual attention modeling," IEEE Trans. Pattern Anal. Mach. Intell., 35(1): 185-207, 2013.

[5] K. Koch, J. McLean, R. Segev, M.A. Freed, M.J. Berry, V.Balasubramanian, P. Sterling, “How much the eye tells the brain,” Curr. Biol., 16(14): 1428-34, 2006.

[6] L. Itti, “Models of bottom-up and top-down visual attention,” Ph.D. thesis, California Inst. of Technology, 2000.

[7] H. Xiaodi, L. Zhang, "Saliency detection: A spectral residual approach." in Proc. 2007 IEEE Conference on Computer Vision and Pattern Recognition: 1-8, 2007.

[8] B.W. Tatler, “The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions,” J. Vis., 14: 1-17, 2007.

[9] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell,“SUN: A bayesian framework for saliency using natural statistics,” J. Vis., 8(32): 1-20, 2008.

[10] N.D.B. Bruce, J.K. Tsotsos, “Saliency based on information maximization,” in Proc. Advances in Neural Information Processing Systems, 2005.

[11] J. Harel, C. Koch, P. Perona, “Graph-based visual saliency,” in Proc. Advances in Neural Information Processing Systems, 19: 545-552, 2007.

[12] C. Siagian, L. Itti, “Rapid biologically-inspired scene classification using features shared with visual attention,” IEEE Trans. Pattern Anal. Mach. Intell., 29(2): 300-312, 2007.

[13] G. Li, Y. Yizhou, "Visual saliency based on multiscale deep features," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 5455-5463, 2015.

[14] R. Anna, P. Napieralski, “The visual attention saliency map for movie retrospection,” Open Phys. 16(1): 188-192, 2018.

[15] W. Hongfa, X. Zhou, Y. Sun, J. Zhang, Ch. Yan. “Deep fusion based video saliency detection,” J. Visual Commun. Image Represent., 62: 279-285, 2019.

[16] W. Wang, J. Shen, F. Porikli, “Saliency-aware geodesic video object segmentation,” IEEE CVPR: 3395-3402, 2015.

[17] W. Wang, J. Shen, R. Yang, F. Porikli, “Saliency-aware Video Object Segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., 40(1): 20-33, 2018.

[18] S. Meijun, Z. Zhou, D. Zhang, Z. Wang. “Hybrid convolutional neural networks and optical flow for video visual attention prediction,” Multimed. Tool. Appl. 77(22): 29231-29244, 2018.

[19] K. Petros, P. Maragos. “SUSiNet: See, Understand and Summarize it,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[20] C. Tran, “Color-opponent channels,” 2014.

[21] S. Kolkur, et al. “Human skin detection using RGB, HSV and YCbCr color models”, arXiv preprint arXiv:1708.02694, 2017.

[22] K. Petros, A. Katsamanis, P. Maragos. “Predicting eyes’ fixations in movie videos: Visual Saliency experiments on a new eye-tracking database," in Proc. International Conference on Engineering Psychology and Cognitive Ergonomics: 183-194, Springer, Cham, 2014.

[23] Z. Bylinskii, J. Tilke, O. Aude, T. Antonio, D. Frédo, "What do different evaluation metrics tell us about saliency models?," IEEE Trans. Pattern Anal. Mach. Intell., 41(3) :740-757, 2018.

[24] A. Mohammadi Anbaran; P. Torkzadeh; R. Ebrahimpour; N. Bagheri. "Fast and efficient hardware implementation of 2D gabor filter for a biologically-inspired visual processing algorithm," *J. Electr. Comput. Eng. Innovations*, 9(1): 93-102, 2021.

[25] S. Mohseni; G. Ardeshir; N. Zarei. "Facial expression recognition based on anatomical structure of human face," *J. Electr. Comput. Eng. Innovations*, 2(2): 77-83, 2014.

[26] M.R. Pishgoo; M.R. N. Avanaki, R. Ebrahimpour, "The application of multi-layer artificial neural networks in speckle reduction (Methodology)," *J. Electr. Comput. Eng. Innovations)*, 2(1): 37-42, 2014.

[27] J. Khosravi; M. Shams Esfandabadi; R. Ebrahimpour, "Image registration based on sum of square difference cost function," J. Electr. Comput. Eng. Innovations, 6(2): 273-281, 2018.

[28] R. Veale, H. Yoshida, "How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling." Phil. Trans. R. Soc. B 372, 2017.

[29] J. Li, Y. Tian, T. Huang, W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," Int'l J. Comput. Vision, 90: 150-165, 2010.

[30] A. Oliva, A. Torralba, M.S. Castelhano, J.M. Henderson, "Top-down control of visual attention in object detection," in Proc. Int'lConf. Image Processing: 253-256, 2003.

Biographies

Majid Shabani received his M.S. in Artificial Intelligence from computer engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. He is the author of 3 conference publications in his research areas, which include Cognitive and Systems Neuroscience, Human and Machine Vision, and Visual Attention. (e-mail: 97shabani@gmail.com)

Alireza Bossaghzadeh is an assistant Professor at the faculty of computer engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. He obtained a Ph.D. degree in the field of Artificial intelligence from the University of the Basque Country, San Sebastian, Spain. His research interests are Manifold Learning, Graph-based Semi-supervised Learning, and Machine vision.

Reza Ebrahimpour is a professor at the Faculty of computer engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. He obtained a Ph.D. degree in the field of Cognitive Neuroscience from the SCS, IPM, Tehran, Iran in July 2007. Dr. Ebrahimpour is the author or co-author of more than 100 international journal and conference publications in his research areas, which include Cognitive and Systems Neuroscience, Human and Machine Vision, Decision Making and Object Recognition.

How to cite this paper:
M. Shabani, A. Bosaghzadeh, R. Ebrahimpour, "A computational-cognitive model of visual attention in dynamic environments," J. Electr. Comput. Eng. Innovations, 10(1): 163-174, 2022.

DOI: 10.22061/JECEI.2021.7871.443

URL: https://jecei.sru.ac.ir/article_1606.html