Ciência e Natura
ISSN: 0100-8307
cienciaenaturarevista@gmail.com
Universidade Federal de Santa Maria
Brasil

Reza Hashemi, Seyyed Mohammad
A Survey of Visual Attention Models
Ciência e Natura, vol. 37, núm. 6-2, 2015, pp. 297-306
Universidade Federal de Santa Maria
Santa Maria, Brasil

Ciência e Natura, v. 37 Part 2 2015, p. 297–306

ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

A Survey of Visual Attention Models

Seyyed Mohammad Reza Hashemi[1*]

[1]Young Researchers and Elite Club, Qazvin Branch, Islamic Azad University, Qazvin, Iran
smr.hashemi@qiau.ac.ir

Abstract

The present paper surveys visual attention models, showing factors' categorization. It also studies bottom-up models in comparison to top-to-down, spatial models compared to spatial-temporal ones, obvious attention against the hidden one, and space-based models against the object-based ones. It categorizes some challenging model issues, including biological calculations, correlation with the set of eye-movement data, as well as bottom-up and top-to-down topics, explaining each in details.

Keywords: *top-to-down attention, stare control saliency, image interpretation, visual search, glance.*

Recebido: dia/mês/ano Aceito: dia/mês/ano

1 Introduction

A rich flow of visual data enter the eye in each second, for which the immediate processing, without any mechanism to reduce their quantity is wrong and very difficult. The mechanism, offered in this presentation indicates visual attention, in the core of which there is a mechanism of selection as well as the concept of connection. In people, attention is easily paid by the retina which has a fully-fledged central gap with high and a margin with low clarity, whereas visual attention directs this anatomic structure to the important parts of the scene so that more details of the information are collected.

In recent decades many aspects of science have attempted to answer this question. Psychologists have studied the behavioral correlation of visual attention such as change blindness, inattention blindness, and attention blindness. Neuron physiologists have shown how neurons adjust in order to show the objects better. Experts of neuroscience have made a model of real neuron networks which simulates and explains behavioral models. Inspired by these studies, computer optics and robotics try to confront the intrinsic issue of calculations' complexity, so that they could make systems which work straightaway. Although there are currently many models, mentioned in the research area above, we limit ourselves to those that can calculate the saliency mappings in each input picture and video. While the term attention, saliency, and stare are often used interchangeably, each have an accurate description, which can be stated as below:

Attention is a general concept, including all factors that affect the selection mechanism, while they are scene-driven, Bottom-Up (BU), or expectation-driven, Top-to-Down (TD). Saliency directly distinguishes some parts of the picture, which could be objects or areas that seem prominent compared to their proximate parts. The term saliency is considered in BU calculations.

Stare is a harmonic movement of the eyes and head, which is often used as an indicator (1) for attention in natural behavior. For instance, while a human or a robot is moving in the environment and interacts with the surrounding objects, he should control stare to do a task. In this concept, controlling the stare is simultaneously involved with sight, factor, and attention to perform the required sensorimotor harmony (2) for the behavior.

2. Categorization of the Factors

We introduce the work by introducing 13 factors (f_(1..13)), later to be used in the models' categorization. These factors possess calculative and behavioral studies of behavior in their roots. Some describe the factors (f_1,2,3,f_(8..11)), whereas the others (f_(4..7),f_12,13) are not directly depended but are of the same account as those, determining the area of different models' usage.

2.1. Bottom-Up Models against Top-to-Down Models

A major significance of models is whether they are based on BU (f1) or TD (f2) effects.

BU indicators (1) are chiefly based on the features of a visual scene (Vector 2 stimulus) whereas the TD ones (Vector 3's destination) are determined by identifying some phenomena such as knowledge, expectation, award, and current goals.

Outstanding areas, which attract our attention, in a BU concept should be distinct enough from surrounding features. This attention mechanism is also called exterior, automatic, reactive issue or output indicators. BU attentions are quick, impulsive, and more similar to Feed-Forward. A primary example of BU attention is looking at a scene with only a vertical strip among many horizontal ones, in which the attention is immediately attracted to the vertical strip. While many models are put in this category, they can explain a quantitative fraction of eye movement, because most stares are the vector's responsibility.

On the other hand, TD attentions are the responsibility of the vector and the closed ring (1). One of the most popular examples of TD attention has been presented by Yarbus in 1967,

Recebido: dia/mês/ano Aceito: dia/mês/ano

who showed that eye movement depend on the current task or the experiences below: Some people were asked to watch a scene (a room with a family and an unexpected visitor, entering the room) with different conditions in order to answer the questions, concerning "the estimation of family's material circumstances" and "the age of the people", or to simply review the scene freely (Fig. 1-1). Eye movement for each of the items varied significantly.

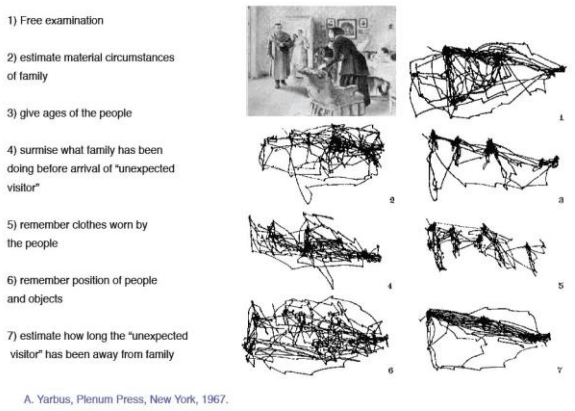

Fig. 1-1 Human eye movement in different circumstances

Models were studied in three chief sources of TD effects in answering this question: How should we decide where to look? Some models arrange visual attention as what attracts our attention to the feature of an object, which we are observing. Others, however, search picture 1's content role in order to limit the areas we are watching.

2.1.1. Features of Objects

There is a significant amount of evidence for goal-based attention in real world searches. Assuming a search in a scene in which the goal is a red element, the attention is quickly drawn to the red element in it. Compare this with a complicated goal object such as a pedestrian in a natural scene; though it is difficult to determine the goal in the latter, there are some features to attract the visual attention (vertical shape, the head as round and the body as straight).

Guided search theory suggests that attention could be biased by modulating the relative gains towards the noticeable goals, based on which varying feature can be involved in the attention. Returning to the previous example, while we are watching a red object an excessive gain could be given to the red color.

From optimum accumulation of signs for goal detection, Navalpakham and Itti resulted in maximizing the rate of the goal's signal to noise in relation to the background. In some earlier studies prior to gathering the mappings for an object location, a weighted function has been performed, which is based on measuring the object's uniqueness to each mapping.

Butko and Movellan modeled an object search, based on some principles of visual search, as stated by Najmenik and Geisler, in a relatively-observable framework for recognition and tracking; however, they did not perform it on eye stare during the face search.

Borji used evolutionary algorithms to search in a space of primary saliency model parameters, in order to find the destination.

The abovementioned studies, carried out on the roles of object features in visual search, are closely connected to the methods of object detection in computer vision. Some approaches of object detection have high detection accuracy for many objects such as cars, people, and faces. On the contrary, recognition models are some purely-calculative approaches. Research on how these two areas are related will most likely result in mutual gains for both.

2.1.2. Scene Context

During a short depiction of an image (about 80 milliseconds), an observer could report the essential features of a scene. Such a very rough depiction, also known as Gist, does not involve numerous details about unique objects, but can contain enough information for the distinction of the scene (for instance interior scene against the exterior one).

It is important to keep in mind that Gist does not necessarily do the semantic categorization of the scene. Semantic connection among the objects in

Recebido: dia/mês/ano Aceito: dia/mês/ano

the scene (e.g. a computer is often placed on a table) or the surrounded environment signs, are shown to be capable of playing a significant role in guiding eye movement.

There have been many models for Gist, by means of different kinds of low-level features. Oliva and Torralba have calculated the range of Fourier transform of a window on other windows, which are not placed on each other, in an image. Afterwards they performed the Principal Component Analysis (PCA) as well as Independent Component Analysis (ICA) to reduce the features' size. Renninger and Malik executed Gabor Filters on an input image, then to extract 100 general texton, selected from a training set, by means of K-Mean Clustering. Their Gist vector is a histogram of these general textons.

Siagin and Itti used biological surrounding center features of direction, color, and channel intensity for Gist modeling. Torralba used regulated wavelet analysis on 6 directions and 4 scales. In order to extract the Gist, a vector is calculated by the average of each output filter on 4×4 cells. Like the previous method, he executed PCA on 384-dimension result vectors so that he could obtain an 80-dimensional vector.

Gist presentation has become increasingly popular in computer vision, because they still provide rich general information for many usages such as search in scene datasets with a vast scale, which still exists.

2.1.3. Requested Task

Tasks have a strong influence on attention development. There have been numerous claims, saying that visual scenes in a need-based method have been interpreted in order to do the requested task. Hayhoe and Ballard showed that when we face a complicated task, there is a strong connection between image recognition and eye movement.

An individual may often keep an algorithm for eye movement in his mind. For instance in a "block copying" task, where people should collect the elements to build the blocks, the observant algorithm has become obvious to complete the task by the eye movement pattern. People first choose a destination block in a model for reviewing the location block, then to stabilize the task space to the new block location in the related place.

Public opinion says that BU and TD attentions are mixed together to direct us to attention behaviors. An integration method should be capable of determining when and how to deal with a TD visual element or omit it as a BU significant sign. Recently a Bayesian approach has been proposed, which considers the optimum accumulation of prize as a TD sign and contrast or orientation as a BU one in humans. Navalpakkam and Itti suggested a recognition model for task-based attention, assuming that the algorithm for doing the task is already available.

2.2. Spatial Models against Spatial-Temporal Models

In real world, we face visual information which is always changing from egocentric or dynamic movements of the world. As knowledge has been accumulated from previous time points, visual selection also depends on current scene saliency; therefore, an attention model should be able to receive those areas of the scene, important in spatio-temporal state.

We can distinguish between two types of temporal information modeling in saliency modeling: 1) some Bottom-Up Models use movement channel to receive human stare, drawn to the moving stimulus. Attention Gate Model (AGM) emphasizes temporal response features of attention, the quantity of level description, and timing for human attention to the destination stimulus, which are consecutive. Previous information about image, eye stare, image context in stare, and physical impact along with other sensorial stimuli (e.g. experience from listening) could be well used to predict the next eye movement. Adding a time dimension as well as natural informal tasks cause some complications in a calculative model in predicting stare on the goals.

Recebido: dia/mês/ano Aceito: dia/mês/ano

Appropriate environments for modeling the temporal aspects of visual attention such as games and films are dynamic and informal. Bioman and Irani proposed an approach to detect the video disorder by comparing tissue sections of the activity to a set of trained data from disorder activities. Temporal information is limited to stimuli level and does not include higher recognition functions such as processed elements at attention center or done activities while they are at play. Some methods result in stable and dynamic saliency mappings, proposing methods that combine them. A modeling approach for spatio-temporal attention for videos has been presented by the combination of movement contrast, itself a result of the homography between two images as well as the measured spatial contrast of color histograms.

F3 indicates whether a model only uses spatial or spatio-temporal information to estimate the saliency.

2.3. Overt Attention against Covert Attention

Attention could be distinct in terms of its characteristics as "overt" against "covert". The former is the process of guiding Fovea towards a stimulus, whereas the latter is mentally emphasizing many sensorial stimuli. An example of overt attention is individual's staring while he is speaking but is aware of visual space, outside Fovea's center. Another example is about driving in which while the driver keeps his eyes on the road, simultaneously covertly observes the signs and lights. Current belief states that overt attention is a mechanism to quickly scan the visual field for a significant place. This covert movement is linked with eye movement axes to order an eye movement to that place (overt attention). However it does not completely explain the interaction between overt and covert attentions. For instance, a person might pay attention to the right side of the visual field's margin, actively stopping the eye movement there. Most of these models discover the areas in order to make eye stare and overt direction of a few eye and head movement attractive. The absence of a calculative framework for covert attention could be due to the fact that behavior mechanism and covert attention functions are still unknown. Additionally, it is still not known how to measure covert attention.

2.4. Space-Based Models against Object-Based Models

There is no similar agreement on the scale unit of attention: do we pay attention to spatial places, characteristic, or object? Most psychophysical and neurobiological studies concern space-based attention. Moreover, there is strong evidence of feature-based (discovering an unusual element in one of feature dimensions or balancing the regulation of characteristic curve of selected neurons) as well as object-based (selectivity of attention to either of two objects, for example face against vase illusion) attentions, here. Current belief is that these theories are not incompatible two by two; visual attention could be expanded to each of the candidate units. Here, delivering a subject is not an individual unit of attention. Humans are able to pay attention to some (between four and five) areas at the same time.

In modeling conception, there is a majority of space-based models. Some object-based models have been previously suggested but they lack any interpretation over eye stare. Such weaknesses could vary their rational review. For instance, the limitations of Sun and Fisher Model is in using artificial segmentation of the images, using those pieces of information which might not be available in pre-attention stage (before detecting the object inside the image). Availability of labeled image and vide datasets has made it possible to effectively guide the present research with this respect. The link between object-based and space-based models will remain to be sorted in future. Feature-based models attempt to regulate the characteristics of some feature detectors in order to create a goal object, which has much saliency in the disorganized background. Since there is a close connection between image and object features, this article categorizes feature-based models under object-based ones, as shown in Fig. 7.

The ninth factor, F9, indicates whether a model accords with space-based or object-based

Recebido: dia/mês/ano Aceito: dia/mês/ano

concept, i.e. whether it needs to work with objects instead of spatial areas.

3. Features

Traditionally, in accordance with Feature Integration Theory (FIT) as well as behavioral studies, three features should be considered in calculative models of attention: intensity (or intensity contrast or radiance contrast), color, and orientation. Usually intensity is from the average of three color channels and is processed by center-surround process, itself inspired by Lateral Geniculate Nucleus (LGN) along with V1 Cortex. Color is established as a red-green and blue-yellow channel, inspired by color contrast neurons in V1 Cortex or intermittently by the use of other color spaces such as HSV and Lab. Orientation is often employed as a convolution with oriented Gabor Filters or by means of oriented masks. Movement has firstly been used by neurons, placed in MT and MST that are capable of being selected to the movement orientation. Furthermore, some studies have been added to specific features of attention direction, such as skin color, face, horizontal lines, wavelet, essential part, central bias, deflection, resolution space, light flow, multiple combined orientations (crosses or corners), entropy, ellipse, symmetry, tissue contrast, high saliency average, depth, and local surround-center contrast. While most models used the FIT-proposed features, some approaches have been unified with other features, such as DoG and features from natural scenes, by means of PCA and ICA Algorithms. In order to search the goal, some people have used a descriptive structure of the objects such as local orientations' histograms. F10 Factor categorizes the models according to the feature they use.

4. Stimulus and Task Type

Image stimuli could be distinguished by belonging to each of stable attentions (e.g. array search, stable photos; Factor F4) or dynamic ones (e.g. videos, games; Factor F5). Video games are informal and very dynamic because they do not produce similar stimuli in each performance and have an almost natural presentation. Although they still belong to natural scenes statistically, they do not have a similar noise distribution.

Establishment here is very complicated, controversial, and quite sensitive to the calculations. They also used many recognition behaviors.

The second difference is between combined stimulus (Gabor patches, array search, cartons, image environments, games; Factor F6) and the natural one (almost belonging to it such as scene photos and videos; Factor F7). Since humans live in a dynamic world, videos and informal environments present an appropriate depiction of visual system, compared to stable images.

Another noteworthy field for behavioral study of attention is the factors to establish virtual reality that can be seen in the work by Spragve and Ballard, who executed a real human agent in VR, using Reinforcing Learning (RL) to harmonize the selection activity and visual understanding in a walking task that avoided hitting the obstacles and tried to stay in the side walk and collect the trash.

Factor F8 distinguishes the tasks from each other. Three tasks have been widely considered in attention modeling context up to now: 1) Free view tasks, in which the individuals are assumed to be freely watching the stimulus (here there is not task or question but many interior recognitions are usually employed); 2) Visual Search Tasks, where are asked to find an unrelated element or a particular object in a natural scene; and 3) Informal Tasks, which uses the objects significantly in many real world situations such as driving and football games. These complex tasks involve many sub-tasks such as visual search, object tracking, focus, and segmented attentions.

5. Evaluation Measures

As a result, we should have a model, the output of which is a saliency mapping (S); and we should quantitatively evaluate it by comparing it with eye movement data (or clicking on the position) (G). How do you compare these? We can consider them as a probability distribution

Recebido: dia/mês/ano Aceito: dia/mês/ano

and use Kullback-Leibler or metrics, in accordance with the percentage of measuring the distance between the distributions. Or we could consider S a Bayesian categorization and use the theory of signal detection analysis (the metric of the area beneath ROC Curve, which the characteristic curve of system performance) to analyze the task of this categorization. We also could consider that S and G are random variants and use correlataion coefficient or can use the normalized saliency of paths' scan to measure their statistical relations. Another road is to consider G as a sequence of eye stare (scanpath) and compare this sequence with a sequence of the stares, selected by a saliency model, called String-Edit Distance (which is a path to determine the likeliness of two similar strings).

While each model might be measured by each criterion, in Fig. 7 we have listed factors in Factor F12 which have been measured by the authors of each model.

Afterwards, when we use Estimated Saliency Mapping (ESM), we mean saliency mapping of a model as well as Ground-truth Saliency Mapping (GSM), a mapping which is created by combining labeled salient areas by a human observant.

The other evaluation criterion, for attention models are categorized into three classes: 1) point-based, 2) area-based, and 3) mental evaluation. In point-based measuring, salient points from ESM are compared to those with GSM in order to combine eye stare. Area-based measuring are suitable for the evaluation of attention models on salient areas of datasets by comparing ESM and labeled salient areas (GSM is interpreted by human mind).

In the following, we focus on the metrics with greater consensus than the literature and provide some signs for others as reference.

Kullback-Leibler Divergence. KL Divergence is usually employed for measuring the distance between two probability distribution, In terms of saliency, it is used to measure the distance between the distribution of saliency in human against random position of the eye, assuming that $t_i = 1..N$, in which N is human eye movement in an experiential period. For a saliency model, EMS in human eye movement is sampled as $x_i, human$ and in a random point as $x_i, random$ (or in a small vicinity is averaged). Afterwards, saliency intensity in the sampled areas is normalized in [0, 1] range. The histogram of these amounts in q areas covers [0, 1] range, calculated throughout quick eye movements. H_k and R_k are a fraction of the points in K areas for salient and random points. Finally, the difference between these histograms with KL Divergence is as follows:

$$KL = \frac{1}{2}\sum_{k=1}^{q}\left(H_k log\frac{H_k}{R_k} + R_k log\frac{R_k}{H_k}\right)$$

Models that can predict human eye stare better, show higher KL Divergence, because the observant usually stares at a quantitative part of the areas (minority) with the highest model response, whereas it avoids the majority of the areas with lowest model response. The advantage of KL Divergence to other samples is that 1) other measures usually calculate the transmission to the right side of H_k, the histogram related to HistogramR_k, whereas KL is sensitive to any kind of difference between the histograms; and 2) KL has no positive influence on repeated parameters such executing each uniform and continuous non-linear on ESM rates of mapping S. One of the disadvantages of KL is that it has no definition to the limit beyond it. As two histograms get completely separated from each other, KL Divergence gains an unlimited approach.

Normalized Saliency Scanpath. It is defined as the response rate at human eye position. (x_h, y_h) is in a normalized ESM model that has zero medium and a standard deviation unit of $NSS = \frac{1}{\sigma^2}(S(x_h, y_h) - \mu_s)$. Similar to percentage calculation, NSS is measured once for each eye movement and subsequently the medium error and standard error are calculated among a set of NSS advantages. NSS = 1 indicates that people's eye movement are set in an area, whose predicted density is higher than the average limit of standard deviation. What is more, $NSS \leq 0$ indicates that the model does

Recebido: dia/mês/ano Aceito: dia/mês/ano

not do an action better than selecting a random situation in a mapping. Unlike KL and in terms of percentage, NSS is not stable in relation to repeated parameter making.

The Area under the Curve (AUC). AUC is the area under the index curve, receiving ROC Factor. As a measure of the best in the society, ROC is being used to evaluate a Bayesian Categorization System with a threshold variant (which is usually used to categorize between two saliency-like methods against the random ones). By using these measures, ESM Models behave as a Bayesian Categorization on each pixel of the image. Pixels with high saliency in relation to a threshold are categorized as a stare while the remaining pixels are grouped as non-stare.

Human stare has then been used as a basis. By using the amounts of different threshold, ROC curve is drawn as the rate of False Positive against True Positive and the areas under the curve show how saliency mappings predict real focus of human eye well. Perfect prediction is equal to 1 point. This characteristic measure wants the stable changes in the area under the curve ROC which does not alter when each constant increasing function is performed on saliency measure.

Linear Correlation Coefficient (CC). This measure is widely used to compare the relation between two images with their uses such as image record, object detection, and dissimilarity measuring. Linear Correlation Coefficient measures the linear connection resistance between two variants:

$$\mathrm{CC}(\mathrm{G},\mathrm{S}) = \frac{\sum_{x,y}(G(x,y)-\mu_G).(S(x,y)-\mu_S)}{\sqrt{\sigma_G^2.\sigma_S^2}}$$

While G and S respectively show GSM and ESM (stare mapping is a mapping with 1s in stare areas and is usually convoluted with a Gaussian), μ and σ^2 are the median and variance of the amounts in these mappings. A significant advantage of CC is the comparison mass of the two variants by providing a simple numerical amount between -1 and +1. When the correlation is near -1 and +1, there is almost a completely linear relation between the two variants.

String Editing Distance. In order to compare the selected noteworthy areas by saliency model (mROI) with the human notable areas (hROI) by means of this measure, saliency mappings and human eye movements are firstly clustered to some parts of the areas. Afterwards ROIs are sorted by the rate, attributed by saliency algorithm or transient sorting of human focus in the scanpath. The results are strings of sorted points such as $string_h = \text{"abcfeffgcd"}$ and $string_s = \text{"afbffdcdf"}$.Sting editing is like S_s Index, which is defined by an optimizing algorithm with the assigned cost unit for three different operations: discovery, insertion, and exchange. Finally the similarity chain between the two strings is defined as =. For instance the string, presented above, is similar to:

6. Datasets

Here, there are many datasets of stable images (for stable attention studies) as well as videos (for dynamic attention studies). Fig. 7 lists F13 as some datasets. We indicate only the datasets, which are used chiefly for measuring and evaluating attention models. All the same, there are other tasks that collect the data for specific goals (e.g. driving, sandwich making, and copying a block). Figs 1.2 and 1.3 illustrate a summary of image and video datasets for eye movements (for a limited amount of labeled salient areas, available). Also researchers have used mouse tracking to estimate attention. Although such type of data is noisy, some recent results prove it to be a rationally-good estimate. For instance Scheier and Egner showed that the pattern of mouse movement is close to the pattern of eye tracking. A web-based mouse tracking work has been established in TCTS Lab.

Recebido: dia/mês/ano Aceito: dia/mês/ano

Study	Subjects	Dataset Size	Resolution	Viewing distance [cm]	Presentation time (s)	Description
Kienzle et al. [165]	14	200	1024 x 768	60	3	8-bit grayscale stimuli presented on a 19-inch Iiyama CRT at full screen size corresponding to 37° × 27° of visual angle.
Einhauser et al. [84]	7	54	640 x 480	50	-	Overall 32,225 fixations with average fixation duration as 370±293 ms and 11.9 fixations per image. The average distance of subsequent fixation points on the screen is 127 pixels(19°). Authors restricted their analysis to 76° ×55° regions which accounts for 98% (29,725) of all fixations. Stimuli was presented using NEC LT 157 projector at resolution 1024 × 768 at 60Hz on average spanned 133 × 100cm, corresponding to 37° × 27° of visual angle.
Ouerhani et al. [210]	6	-	640 x 480	70	5	Age range [24-34], with normal or corrected-to-normal acuity as well as normal color vision. Stimulus presented on a 19" monitor subtending 29° × 22°. Task was "just look at the image". Eyetracker: EyeLink, SensoMotoric Instruments GmbH. Recording at 250Hz, accuracy 0.5°– 1° accuracy with a 3×3 point grid calibration sequence.
Bruce and Tsotsos [144]	20	120	681 x 511	75	4	Images [indoor and outdoor] were presented at random with 2 s gray mask in between on a 21-inch CRT monitor. The eye tracking apparatus consisted of an ERICA workstation including Hitachi CCD camera with an IR emitting LED. Stimuli were color images and task was free viewing. Link: www-sop.inria.fr/members/Neil.Bruce
Stark and Choi [211]	7	15	-	40	4	Bright Purkinje reflection captured by a video camera. Stimulus size was 15 × 20cm yielding to 21° × 29° with the 0.5-1 degree accuracy. Images were terrain photographs, landscapes and paintings. Task was free viewing.
Chikkerur et al. [154]	8	220	640 x 480	70	5	Scenes contained cars (4.6 ± 3.8) and pedestrians (2.1 ± 2.2)- visual angle: 16 × 12. Subjects were asked to count the number of cars or pedestrians. Using an ETL 400 ISCAN, table-mounted video-based eye tracker at 240 HZ and accuracy of 0.5°. [age:18-35]. Images were 100 from x and 120 from LabelMe. Link: http://www.sharat.org/
Torralba et al. [92]	24	36	15.8 x 11.9	-	-	In people search task, 14 stimuli out of 36 contained no people and 22 included 1-6 people. The same set [36 indoor] images was used for painting search (17 images without any paintings and rest with 1-6 paintings) and for mug search (half without and half with 1-6 mugs). Eyetracking was performed by a Generation 5.5 SRI Dual Purkinje Image Eyetracker, sampling at 1000Hz. Color photos displayed on a NEC Multisync P750 monitor (143Hz refresh). Mean target size was 1.05%(1.24%) of the image size for people, 7.3%(7.6%) for painting and 0.5% (0.4%) for mugs. Link: http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/
Judd et al. [166]	15	1003	Various	48	3	Images were collected from Flicker creative commons and LabelMe datasets. The longest dimension was 1024 with other ranging from 405 to 1024. There were 779 landscape images and 228 portrait images. Images were freely viewed with 1 sec gray screen between each two. Camera was recalibrated after every 50 images. First fixation was discarded. Age range: 18-35. Link: http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html
Cerf et al. [167]	7	250	1024 x 768	80	-	Eye position of subjects were acquired at 1000Hz using an Eyelink 1000 (SR Research, Osgoode, Canada). The task had three phases: 1) free viewing, 2) searching for face, an object, banana, cell phone, toy car, etc shown by a probe image, and 3) 100 image recognition memory task where subjects had to answer with y/n whether they had seen the image before. Stimuli subtended 28° × 21° of visual angle. Link: http://www.ifadv.com/
Peters et al. [134]	12	100/ class	-	75	-	ISCAN Inc eye tracker was used to sample eye movements at 120Hz. Age range: 18-25; four did free-viewing over [outdoor photos, overhead satellite imagery, and fractals]. Another 4 did free-viewing over involving Gabor snakes and Gabor arrays. Seven subjects did a contour detection task. Resolution was 1000 × 1000 to 1536 × 1024 subtending a visual angle of 15.8° × 15.8° to 16.2° × 25°. Link: http://lab.usc.edu
Reinagel and Zador [212]	5	77	640 x 480	79	10	Images were 69 nature scenes, 38 man-made objects such as buildings, 17 animals or humans and 8 synthetics. An RK-416 Infrared Pupil Tracking System and a 21-inch monitor was used. The whole image subtended 28° × 21° of visual angle. Subjects were instructed to "Study the images". Estimated tracking error was 0.5°. Link: http://zadorlab.cshl.edu/
Hwang and Pomplun [87]	30	160	1280 x 1024	-	10	Age range: 19-40. Stimuli were 160 photographs (1280 × 1024) real-world scenes including landscapes, home interiors, and city scenes and covered 20° × 20° of visual angle. An SR research EyeLink II system. Stimuli presented on 19-inch Dell P992 monitor (85Hz refresh rate), the whole image subtended 28° × 21°. Link: http://www.cs.umb.edu/~marc/
Kootstra et al., [136]	31	99	1024 x 768	70	-	Eyelink head-mounted eye tracking (SR research) was used and was recalibrated before each session. Age range: 17-32. Task was free viewing. Stimuli were: 12 Animals, 12 Automan, 16 Buildings, 20 flowers, 41 natural scenes and were shown on a 18-inch CRT monitor (36 × 27 cm). Link: http://www.csc.kth.se/~kootstra/
Tatler [123]	14	48	800 x 600	60	-	Eyelink-I eye tracker was used. Subjetcs had normal or corrected to normal vision with age range 17-32. Image subtended 30° × 22° and were presented on a 17-inch SVGA color monitor (74 Hz refresh). Task was free viewing. Link: http://www.scienceresearchlab.org/
Engmann et al., [182]	8	90	1280 x 1024	85	-	Subjetcs had normal or corrected-to-normal vision and normal color vision with age range 20-27 (avg: 22.3). Stimuli were presented on a 19.7" EIZO FlexScan F77S CRT monitor (100 Hz refresh). Natural scenes selected from the Zurich natural image database (Einhauser et al. [99]) which only rarely contain isolated nameable objects or man-made artifacts at resolution 2048 * 1536. Image subtended 26° × 18° 17-inch SVGA color monitor. Task was free viewing. Eye tracker was Eyelink-2000 (SR Research Ltd. Canada) with 13 point calibration.
Engelke et al. [213]	30	7	512 x 512	60	8	Images were 4 human faces ("Barbara"), 1 "Glodhill" face (gorilla) and 1 "Peppers" images. Eye tracker was EyeTech TM3 and task was free viewing. Each image was presented for 8 sec with a gray screen with central fixation in between.
Le Meur et al. [41]	40	46	800 x 600	*	-	Stimuli were 46 degraded versions of 10 color images using spatial filtering. Task was free viewing. Eye tracker was made by Cambridge Research Corporation. Viewing distance was four times the TV monitor height. Link: http://www.irisa.fr/temics/staff/lemeur
Ehinger et al. [87]	14	912	800 x 600	75	15	Stimuli were color images (half with a pedestrian) with resolution 800 × 600 and were shown on a 21-inch CRT monitor with resolution 1024 × 768 and refresh rate 100Hz. A 240 Hz ISCAN RK-464 video-based eye tracker was used for recording. The task was to decide wether a pedestrain is in the scene or not. Link: http://cvcl.mit.edu/searchmodels/
Rajashekar et al. [174]	29	101	1042 x 768	134	-	Subjects were 18 males, 11 females with mean age of 27. Eye tracker was made by Image Systems Corp, MN. Grayscale images were shown on a 21-inch grayscale gamma corrected monitor with resolution 1024 × 768. The task was free viewing. Link: http://live.ece.utexas.edu/research/doves/

Fig.1-2 Datasets of the images and the conducted method

Dataset	Features	Feature Value
CRCNS - ORIG [145]	C	50 clips (0.06-1:30 min each), ~25 min total, ~68B for 46K frames
	S	8 (3 female, 5 male) subjects with normal corrected vision, Ages 23-32, From mixed ethnicities
	T	"Follow main actors and actions, try to understand overall what happens in each clip."
	ST	Complex video stimuli involving TV programs, outdoor sceces, video games Outdoor day & night, parks, crowds, rooftop bar, etc.
	D	ISCAN RK-464 eye tracker, 240 HZ recording, 9 point calibration after every 5 clips. 640× 480 resolution at 60.27HZ doublescan, 33.185ms/movie frame, (x,y) of each saccade
	L	http://crcns.org/data-sets/eye/eye-1
CRCNS - MTV [145]	C	50 video clips (4-7 subjects on each video clip)
	S	8 subjects different from subjects of CRCNS
	D	This dataset was created by cutting video clips of CRCNS into 1-3s "clipets" and reassembling those clippets in random order. Other aspects were the same as the original dataset.
	L	http://crcns.org/data-sets/eye/eye-1
Jia Li et al. [133]	C	431 videos with total length of 7.5 hours, 764,806 frames in total with 62,356 key frames
	S	23 (17 male and 4 female) subjects with age range between 21-37
	ST	6 genres: documentary, ad, cartoon, news, movie and surveillance
	D	10-23 subjects per each clip were assigned to manually label the salient regions with one or multiple rectangles from key frames. Drawback with this dataset is rectangular labeling but this may be resolved with segmentation, inefficiney to evaluate whatever
	L	http://www.jdl.ac.cn/user/jiali/
Peters and Itti [101]	C	24 game-play sessions, ~185 GB for 216K frames, 8,449 saccades of amplitude 2o or more
	S	5(3 male, 2 female) subjetcs with normal corrected vision
	T	"Play 4 or 5 five-minute segments of the Nintendo GameCube games"
	ST	Games include Mario Kart, Wave Race, Super Mario Sunshine, Hulk and Pac Man World.
	D	Subjects were seated viewing distance of 80 cm (28° × 21° usable field of view) Stimuli were presented on a 22" computer monitor (LaCie Corp; 640 × 480, 75 HZ refresh, mean screen luminance 30cd/m2, room luminance 4 cd/m2) ISCAN RK-464 eye tracker, 240 HZ recording. 9 point calibration after before game segment Frames were grabbed using a dual-CPU Linux computer with SCHED FIFO schedulingto ensure microsecond accurate timing.
	L	http://lab.usc.edu/njpeters/
Shic and Scassellati [74]	C	2 clips, 10, young adults, normal and mildly mentally retarded
	T	"One minute long clips from back and white movie "Who's afraid of Virginia Woolf"
	D	A head mounted eye-tracker (ISCAN Inc.) was used. The eye tracker employs dark pupil- corneal reflection video-oculography and had accuracy within ±0.3o over a horizontal and range of ±20o, with a sampling rate of 60 Hz. The subjects sat 63.5 cm from the 48.3 cm screen on which the movie was shown at a resolution of 640 × 480 pixels.
	L	http://sites.google.com/site/fredshic/home
Marat et al. [49]	C	53 short video clips (25 fps, 720 × 576 pixels), 1700 frames
	S	15 (3f,12m) subjects with age range 23-40 and had normal or corrected to normal vision
	ST	Each clip ~ 1-3sec long, 324 clip snippets. There was not a particular task or question. TV shows, TV news, animated movies, commercials, sport and music. Indoor, out-door, day-time, night-time)
	D	The clip snippets were strung to form 20 clips of 30 seconds (30.20 × 0.61). Eye positions were recorded at 500 Hz (20 eye positions per frame for two eyes) using a Eyelink II (SR Research). Participants were positioned with their chin supported on a 21" color monitor (75 HZ) at a viewing distance of 57cm (40° × 30° usable field of view). A calibration was carried out at every five stimuli and a control drift was done before each stimuli.
	L	http://start1g.ovh.net/~qgsmabaq/sophie/index.php
Le Meur et al. [138]	C	7 clips (25 Hz, 352 × 288 pixels), 2451 frames, Each clip ~ 4.5-33.8 sec long
	S	17-27 subject for different clips with normal or corrected to normal vision
	T	Free viewing
	ST	Faces, sporting events, audiencesm, landscape, logos, incrustations, low and high satiotemporal
	D	Dual-Purkinje eye tracker from Cambridge Research Corporation. Sampling frequency was 50Hz. CRT display 800 × 600 pixels, 25° × 27°. Distance to screen was 81 cm.
	L	http://www.irisa.fr/temics/staff/lemeur

Fig. 1-3 Datasets of the images and performed features

7. Results

In this article we discussed the improvements of attention, while focusing on saliency models, which have been recently carried out, and

Recebido: dia/mês/ano Aceito: dia/mês/ano

studied Bottom-Up models against the Top-to-Down ones, Spatial Models against Spatio-Temporal ones, Overt Attention against Covert Attention, and Space-Based Models against the Object-Based ones, showing that there is a good number of technical usages which we can use. A promising path for future researches is the development of models, which consider the duties, based on the expected task, especially in informal, complicated, and dynamic environments. In addition, there has not been an essential calculation yet to understand covert and overt attentions, which should be clarified in future.

Reference

[1] Ali Borji, "State-of-the-Art in Visual Attention Modeling," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 35, NO. 1, JANUARY 2013.

[2] L. Itti, "Models of Bottom-Up and Top-Down Visual Attention," PhD thesis, California Inst. of Technology, 2000.

[3] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[4] Y. Zhai and M. Shah, "Visual Attention Detection in Video Sequences Using Spatiotemporal Cues," Proc. ACM Int'l Conf. Multimedia, 2006.

[5] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," J. Vision, vol. 8, no. 32, pp. 1-20, 2008.

[6] L. Itti, "Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes," Visual Cognition, vol. 12, no. 6, pp. 1093-1123, 2005.

[7] R. Rao, "Bayesian Inference and Attentional Modulation in the Visual Cortex," NeuroReport, vol. 16, no. 16, pp. 1843-1848, 2005

[8] Suh, B., Ling, H., Bederson, B. B., and Jacobs, D.W., "Automatic thumbnail cropping and its effectiveness,"

in [ACM Symposium on User Interface Software and Technology], 95–104 (2003).

[9] Zhang, M., Zhang, L., Sun, Y., Feng, L., and Ma, W., "Auto cropping for digital photographs," in [IEEE

Intl. Conf. on Multimedia and Expo], (2005).

[10] Stentiford, F. W. M., "Attention based auto image cropping," in [Workshop on Computational Attention

and Applications, ICVS], (2007).

[11] Ke, Y., Tang, X., and Jing, F., "The design of high-level features for photo quality assessment," in [IEEE

Conf. on Computer Vision and Pattern Recognition], 419–426 (2006).

[12] SMR. Hashemi, M. Zangian, M. Shakeri, and M. Faridpoor, "Survey Article about Image Fuzzy Processing Algorithms." The Journal of Mathematics and Computer Science, Vol 13, Issue 1 2014, pp 26-40

[13] SMR. Hashemi, "Review of algorithms changing image size." Cumhuriyet Science Journal, Vol. 36, No: 3 Special Issue (2015)

[14] SMR. Hashemi, M. Kalantari, and M. Zangian, "Giving a New Method for Face Recognition Using Neural Networks.", International Journal of Mechatronics, Electrical and Computer Technology Vol. 4(11), A pr, 2014, pp. 744-761, ISSN: 2305-0543

[15] SMR. Hashemi, A. Broumandnia, " A Review of Attention Models in Image Protrusion and Object Detection." The Journal of Mathematics and Computer Science, Vol 15, Issue 4 2015, pp 273-283

[16] M. Mohammadpour, SMR. Hashemi and A. Broumandnia, "Improve Image Re-targeting Algorithm Using Markov Random Field", The second international conference on Pattern Recognition and Image Analysis, 2015

[17] SMR. Hashemi, A. Broumandnia, " A New Method for Image Resizing Algorithm via Object Detection ." International Journal of Mechatronics, Electrical and Computer Technology, Vol 5, Issue 16 2015.

Recebido: dia/mês/ano Aceito: dia/mês/ano