Contents list available at JMCS

Journal of Mathematics and Computer Science

Journal Homepage: www.tjmcs.com

A Review of Attention Models in Image Protrusion and Object Detection

Seyyed Mohammad Reza Hashemi[1], Ali Broumandnia[2]

[1]Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
smr.hashemi@qiau.ac.ir – (corresponding author)

[2]Faculty of Computer and Information Technology Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

Article history:
Received March 2015
Accepted July 2015
Available online July 2015

Abstract

Modelling in visual attention especially the stimulus-driven one, i.e. saliency-based attention, has been a very active research field during the recent 25 years. There are many attention models which, apart from being in other aspects, have been offered in successful functions of computer vision, moving robots, and cognitive systems. The present article surveys the primary concepts of visual attention, implemented in cognitive, Bayesian network, decision theories, and information theory in a computational perspective. It will demonstrate a categorization that provides a critical comparison of the approaches as well as their abilities and results. Specifically, the article formulates the criteria, derived from computational behaviors and studies in order to compare the quality of visual attention models.

Keywords: visual attention, bottom-up attention, top-down attention, saliency, eye movement, regions of interest, visual search.

1. Introduction

One of the clear features of humans is their effectiveness in environments where they receive much sensory information. Sight is the most important sense on which people depend; that is why this sense has attracted the most number of studies in machine sight as well as artificial intelligence. Despite wide researches in machine and robotic sight, a good number of sensory-motor actions, done easily by the humans, have not been solved yet. Particularly, it is a matter of great interest to design learning algorithms with high accuracy and low computational complexity which enables automatic dynamic robots to function in visual interactive environments. In comparison to environments with controlled sight, more often than not used in laboratories, learning visual behavior in uncontrolled and general environments are quite harder. Instances of visual learning usages are sight-oriented guidance, positioning via visual information, and obtaining and moving objects.This recent tendency in robotics heads towards the development of robots which can automatically function in unknown and random visual environments which appropriates and necessitates interactive and linear methods for visual representations and control. Such dynamic methods result in flexible solutions with few complexities and low computational costs. To be able to function in visual environments, a robotic factor should manage to have its physical actions correspond to its cognitive visual space.

Thisfeature, visual-motor harmony, is called intention-based sight or sight for action. Unlike machine visual solutions, which often presuppose static and pre-defined representations in the factor's mind, necessary representations in intention-based sight are formed from the interaction of the factor with its surrounding environment.In the recent decade, many aspects of science tried to answer this question with psychologists studying the behavioral correlate of visual attention such as change blindness, inattention blindness, and attention blink.

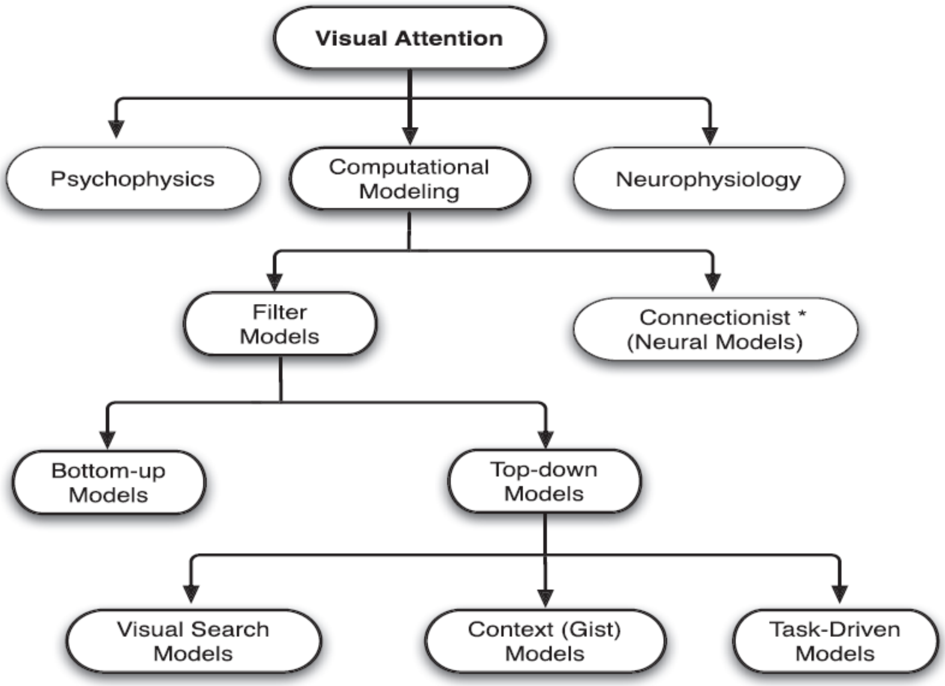


Figure 1: Categorization of attention studies

2.1. Definitions

Although attention, protrusion, and stare are often used interchangeably, each has an accurate meaning of its own, which will be defined below.Attention is a general concept, involving all factors influential on the selection mechanism and are scene-driven (bottom-up) or expectation-driven (top-down).Protrusion directly involves some parts of the picture which could be objects or areas, seeming to be protruding compared to their surrounding parts. This term is considered as bottom-up calculations.Stare is a harmonious movement of eyes and head, which is often used as a proxy for attention in natural behaviors, for example, a man or a robot while in movement and interacting with surrounding objects, should control its stare in order to function. In this concept, to control the stare is simultaneously involved with sight, factor, and attention to carry out the required sensorimotor harmony.

2.2. Starting Point

The basis of all attention models is "Feature Integration Theory" of Treisman and Gelade who claimed that visual features are important, working on how these features must be combined to guide human attention to following and connection-oriented tasks. Afterwards Koch and Ullman, suggested a Feed-Forward Model that combined these features and was an introduction for the concept of Saliency Map, itself a topography to show the protrusions of the areas in a scene. Moreover, they introduced a neural network which selected the most protruded areas but had an inhibition, preventing the attention to be drawn to the next most protruded area. The first complete implementation and revision of Koch and

Ullman’s model was done by Itti (Fig 1-2) who applied it on a combination of natural scenes. Various approaches with different presuppositions has been suggested for attention models, evaluated by means of different sets of data.

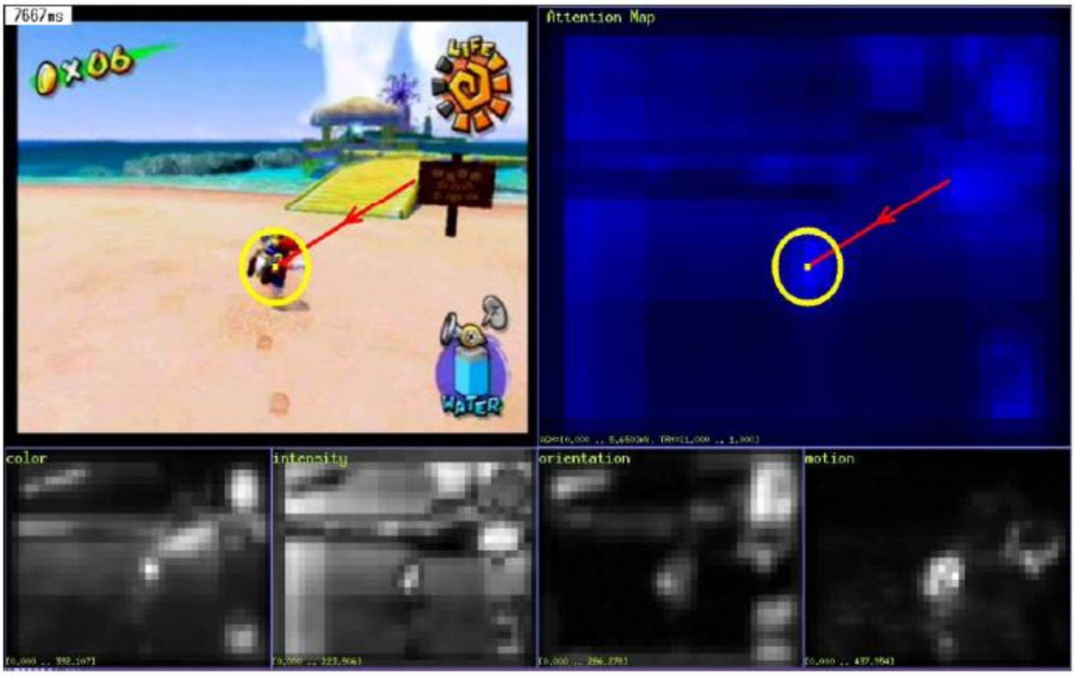


Figure: 1-2 developed tool in iLab

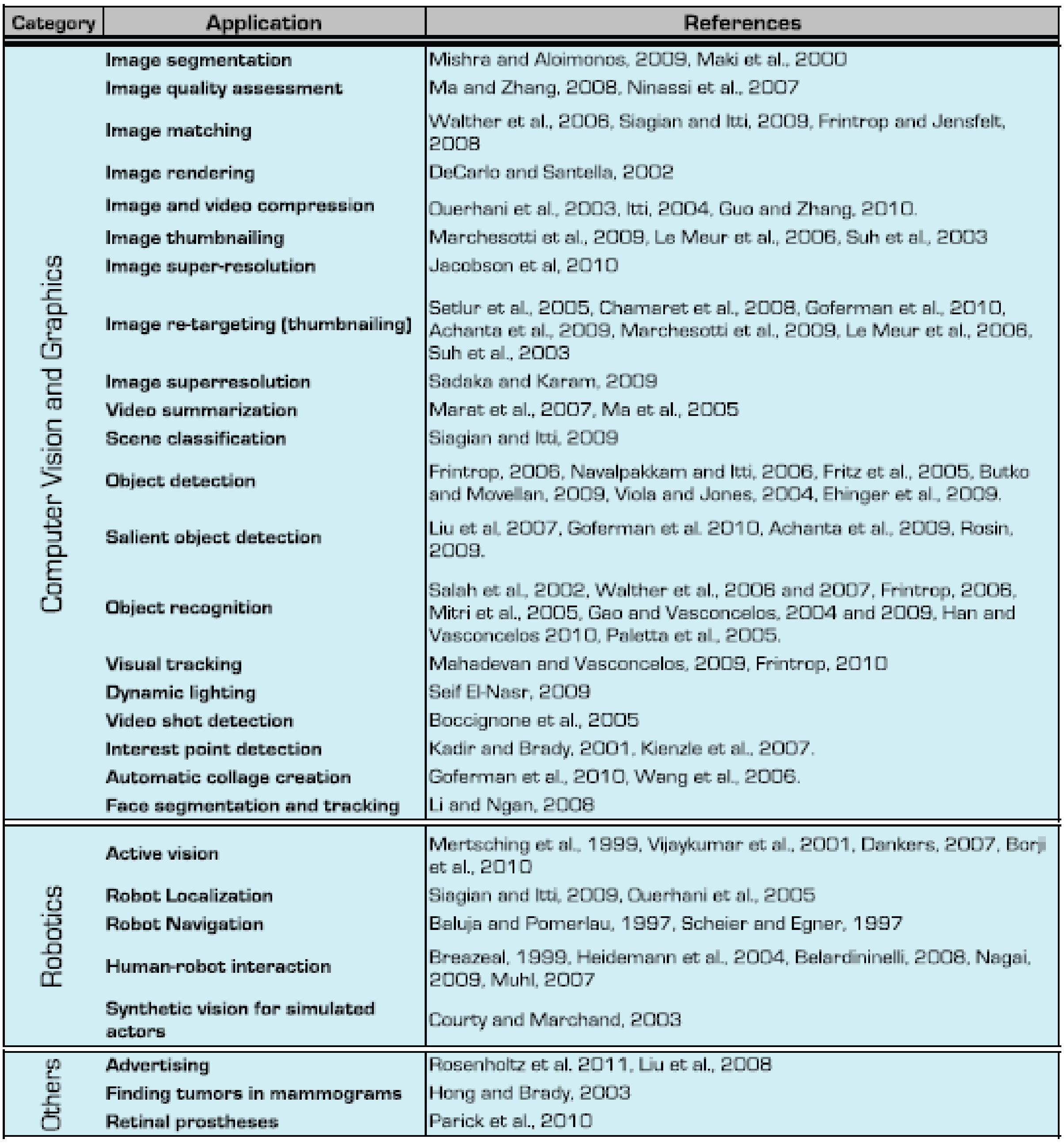
Main importance of attention modelling is how, when, and why we select the behaviors, related to image areas. There are many definitions and computational aspects due to these factors. A general approach is inspired by human’s agency as well as anatomy of sight system, which had been developed a lot. Alternatively some studies hypothesize that which function of visual attention can be used and formulated in a computational framework. For example some claim that visual attention is attracted to places with suitable information, areas of the image that are wonderful, or areas which maximize the reward during carrying out the task.

2.3. Scientific Bases

Attention models are usually analyzed before human observers’ eye movement which show the important information, regardless of cognitive process such as reading, visual searching, and scene understanding. Moreover they behave as an agent to change behavior attention. For example in scene understanding and visual search, when the scene is very cluttered and crowded, the stare is longer and fast eye movement is very shorter. Task difficulty (for instance reading to understand against reading in general, or searching for a person in a scene against watching a scene for memory test) obviously affects eye movement behavior. Even though both attention and predicting models of eye movement have often been analyzed by eye data, there are little differences in terms of area, approaches, stimulus, and detail level. Predicting models of eye movement try to understand the mathematics and underpinnings of attention. It should be taken into consideration that eye movement does not always tell the whole story and there are other metrics which can be used for model evaluation, such as accuracy in correct report of a change in an image. Many attention models have been tested for visual seach by means of accurate estimating of reaction time. The focus of this article will be to define attention models. All in all, many technological usages of these models have been developed during recent years, increasing the attention capability in these models.

2.4. Functions

Functions of attention models are put in three categories: visual and graphic, robotic, and the rest, which are demonstrated in Fig. 1-3:



Category	Application	References
Computer Vision and Graphics	Image segmentation	Mishra and Aloimonos, 2009, Maki et al., 2000
	Image quality assessment	Ma and Zhang, 2008, Ninassi et al., 2007
	Image matching	Walther et al., 2006, Siagian and Itti, 2009, Frintrop and Jensfelt, 2008
	Image rendering	DeCarlo and Santella, 2002
	Image and video compression	Ouerhani et al., 2003, Itti, 2004, Guo and Zhang, 2010.
	Image thumbnailing	Marchesotti et al., 2009, Le Meur et al., 2006, Suh et al., 2003
	Image super-resolution	Jacobson et al, 2010
	Image re-targeting (thumbnailing)	Setlur et al., 2005, Chamaret et al., 2008, Goferman et al., 2010, Achanta et al., 2009, Marchesotti et al., 2009, Le Meur et al., 2006, Suh et al., 2003
	Image superresolution	Sadaka and Karam, 2009
	Video summarization	Marat et al., 2007, Ma et al., 2005
	Scene classification	Siagian and Itti, 2009
	Object detection	Frintrop, 2006, Navalpakkam and Itti, 2006, Fritz et al., 2005, Butko and Movellan, 2009, Viola and Jones, 2004, Ehinger et al., 2009.
	Salient object detection	Liu et al, 2007, Goferman et al. 2010, Achanta et al., 2009, Rosin, 2009.
	Object recognition	Salah et al., 2002, Walther et al., 2006 and 2007, Frintrop, 2006, Mitri et al., 2005, Gao and Vasconcelos, 2004 and 2009, Han and Vasconcelos 2010, Paletta et al., 2005.
	Visual tracking	Mahadevan and Vasconcelos, 2009, Frintrop, 2010
	Dynamic lighting	Seif El-Nasr, 2009
	Video shot detection	Boccignone et al., 2005
	Interest point detection	Kadir and Brady, 2001, Kienzle et al., 2007.
	Automatic collage creation	Goferman et al., 2010, Wang et al., 2006.
	Face segmentation and tracking	Li and Ngan, 2008
Robotics	Active vision	Mertsching et al., 1999, Vijaykumar et al., 2001, Dankers, 2007, Borji et al., 2010
	Robot Localization	Siagian and Itti, 2009, Ouerhani et al., 2005
	Robot Navigation	Baluja and Pomerlau, 1997, Scheier and Egner, 1997
	Human-robot interaction	Breazeal, 1999, Heidemann et al., 2004, Belardininelli, 2008, Nagai, 2009, Muhl, 2007
	Synthetic vision for simulated actors	Courty and Marchand, 2003
Others	Advertising	Rosenholtz et al. 2011, Liu et al., 2008
	Finding tumors in mammograms	Hong and Brady, 2003
	Retinal prostheses	Parick et al., 2010

Figure: 1-3 Categorization of attention models

2.5. Description and Organization

It is hard to define attention officially, while it has been universally accepted. All the same, from a computational viewpoint, many attention models can be unified under general issues: Assuming that K subjects are displayed in a set of N images in which I $= \{\boldsymbol{I}_i\}_{i=1}^{N}$. Assuming that $L_i^k = \{p_{ij}^k, t_{ij}^k\}_{j=1}^{n_i^k}$ is a vector of staring $p_{ij}^{k} = (x_{ij}^{k}, y_{ij}^{k})$ and t_{ij}^{k} is the time for the Kth subject in image $\boldsymbol{I}_i$. If the number of stares of this subject in ith image is p_i^k, the aim of attention modeling is to find a function (mapping the stimulus's protrusion) of $f \in \mathcal{F}$ which minimizes the error of predicting the staring eye, i.e. $\sum_{\mathrm{k}=1}^{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{m}(f(I_i^k), L_i^k)$ in which $m \in \mathcal{M}$ is a measure for distance. There is a key point here; the above definition is suitable for bottom-up attention models. It is not necessarily possible for it to include all aspects of visual attention, for example in case of covert attention or top-down factors that cannot be explained by the staring eye. Here we demonstrate a systematic view of major attention models, which we can apply on desired images.

3. Attention Models

The models are explained based on their mechanism of obtaining protrusion. Some models are places in more than one set. Below, we have focused only on these models which are implemented in software way and can process the desired digital images, returning the protrusion mappings. The models are introduced in the order of their history. It should be noted that here we pay attention to protrusion models instead of their approaches, detecting and segmenting the most protruded areas or objects. As long as these models are used as a protrusion operator in the initial levels, their main goal is not to explain attention behavior. All the same some methods are more inspired from the next protruded models. We use the phrase "protrusion detection" that refer to these approaches.

3.1. Cognitive Models

Almost all attention models are inspired from cognitive concepts either directly or indirectly. One which has more dependency on psychology or neurophysiology is discussed below. In his simple models article, Itti used three feature channels of color, intensity, and direction. This model became the essence of other standard models and stimuli for comparison. Moreover, his work is shown in relation to human eye movement in free-view tasks. An input image to Gaussian Pyramids is sampled with each Gaussian surface G, disintegrated to channels for red (R), green (G), blue (B), yellow (Y), intensity (I), and local direction (O_θ). Among these channels, surrounding center of "feature mapping", f_l, is made and normalized for different features, L. In each channel, are recollected along the scale and re-normalization:

$$f_l = \mathcal{N}\left(\sum_{c=2}^{4} \sum_{s=c+3}^{c+4} f_{l,c,s}\right), \forall l \in L_I \cup L_C \cup L_O,$$

$$L_I = \{I\}, L_C = \{RG, BY\}, L_O = \{0°, 45°, 90°, 135°\}$$

These mappings are collected and normalized linearly more than once so that they can give an obvious mapping:

$$C_I = f_I, C_C = \mathcal{N}\left(\sum_{l \in L_C} f_l\right), C_O = \mathcal{N}\left(\sum_{l \in L_O} f_l\right)$$

Finally obvious mappings are combined linearly more than once in order to produce a protrusion mapping: $S = \frac{1}{3} \sum_{k \in \{I,C,O\}} C_k$

Le Meur has suggested an approach for bottom-up protrusion based on the structure of human visual system (HSV). Instensity-sensitive functions, perceptual decomposition, visual masking, and activities of surrounding center are some implemented features in this model. Later and Le Meur expanded this model to a spatial-temporal domain by means of achromatic fusing, full color, and temporal information. In this new model, previous visual features are extracted from visual inputs into many parallel channels. A feature mapping for each channel is provided and afterwards an outstanding protrusion mapping of the combinations of those channels is generated. Navalpakhm and Itti have made visual search as an optimization issue of top-down benefit by maximizing the signal-to-noise rate (SNR) from the destination against distracting factors instead of teaching the simple functions. As a result, they have taught linear volumes for feature combinations via maximizing the rate between destination and distracting factors protrusions. Kootstra et al. have developed three protruded-symmetrical operators, comparing them with human eye detecting data. Their method was based on symmetrical operators, equal from every direction, and symmetrical with Reisfeld operator beam as well as Heidemann color. Kootsrta has developed these operators to multi-scaled models of symmetrical protrusion. Writers have shown that their models effectively work better on symmetric stimuli, compared with the method by Itti et al.

Marat et al have suggested a top-down approach to predict the spatial-temporal protrusion on video stimuli. This model extracts two signals from the video flow, related to parvocellular and magnocellular cells, from the retinal. Two static and dynamic protrusion mappings derive from these signals, combining with a spatial-temporal mapping. Prediction results of this model were better for the first new frame of each clip. Murrar et al introduced a model, based on low-level sight system in three stages: 1) Sight stimuli are followed in accordance to the knowledge about human sight direction (color contrast and light canals) and are processed by a multi-scale analysis. 2) A simulation of the deterrent mechanism in spatial sight cells, normalizing their responses to conflict stimuli, is prepared. And 3) the information is collected in multiple scales by directly carrying out an inverse wavelet transform on measured volumes from membranous output normalization. Cognitive models, in relation to our biological view, have the advantage of visual attention support. These advantages help understanding calculative principles or sensational mechanisms of this process, as well as other dependant complicated processes, such as object detection.

4. Bayesian Models

Bayesian Models have been used to combine sensational evidence with previous limitations. In these models, the previous knowledge (for example sense of content or general view) as well as sensational information (for instance destination features) are combined based on Bayesian Law probabilistically (for example discovery of an outstanding object). Torralba, Oliva, et al. have offered a Bayesian work framework for visual search tasks. Bottom-up protrusion is derived from the formula $\frac{1}{p(f|f_G)}$ in which FG indicates a global feature, summarizing the possibility of the target object's density in the scene based on total analysis of the scene, itself. Ehinger et al. have done the linear gathering of three components (bottom-up protrusion, total view, and object features) to explain people's eye movement when searching in a databank of 900 natural scenes. Itti and Bladi have introduced the amazed stimulus, effectively changing an observer's beliefs. This has been modeled in a Bayesian framework by convergence calculation of KL between future and previous beliefs. This symbol have been applied on both space (amazement occurs when the features of the observed image in a visual situation affects the observant of the neighboring locations) and time (amazement occurs when observed image features in a point of time affects the belief, formed from previous observations).

Zhang et al. has given a definition of recognized protrusion, called SUN (Saliency Using Natural) which considers which visual system tries to optimize when the attention is direct. Resultant model is a Bayesian framework, in which bottom-up protrusion naturally appears as the self-information of visual features and total protrusion (unification of the top-to-down information with bottom-up protrusion) shows up as the opposite of information between image's local features and destination feature searches, once the search is being performed for a specific target. Since this model provides a general framework for many other models, it will be described in details. SUN for bottom-up protrusion is similar to the work by Oliva et al., Torralba, Bruce, and Tsotsos, all of which are based on self-information symbols (local information). All the same, the difference is between flowing image statistics and the natural one. In a nutshell, the motivation factor to use self-information with flowing image statistics is that there is a similar background object which has outstanding features of this background. Since intermittently targets are observed less than the background in a period, rare features are much similar to the shown targets. Assuming that Z indicates a pixel in an image, if C is a point belonging to the target class and L is the point's location, and if F is an image feature of the point, the sz protrusion of the point Z can be defined as $P(C = 1|F = f_z, L = l_z)$ where fz and lz are feature and location of z respectively. Using Bayesian Law and assuming that the feature, position, and the known conditional position is C = 1, the protrusion of a point will be:

$$log\ \mathrm{s_z} = -\ logP(F = f_z) + logP(F = f_z|C = 1) \\ + logP(C = 1|L = l_z).$$

The first item on the right side of self-information is bottom-up protrusion, only dependant on visual features, observed in Z and the second right-side one is the correction logarithm, in which the amounts of the feature in question are harmonious with the target's previous knowledge (i.e. if the target is known to be green, correction logarithm intends to accept bigger amounts for a green point in comparison to a blue one). The third item is the previous location, in which top-to-down knowledge receives the target's location and it is independent from objects' visual features. For example, this item might receive some knowledge about some targets which are often found on the top left corner of an image. Zhang et al have expanded SUN Model to dynamic sensations by introducing temporal filters (different from exponential), applying a generalized Gaussian distribution for each filter's response. This method has been first implemented a databank of spatial-temporal filters on each video frame. Afterwards for each video, the model calculates the features of each point, estimating its bottom-up protrusions. Filters are designed to be effective and similar to human sight system. The distribution of these spatial-temporal features is taught from a set of video shapes of natural environments. Li et al presented a Bayesian multi-tasking learning framework for visual attention in videos. Bottom-up protrusion is modeled by multi-scale wavelet analysis, which was combined with up-to-down parts, taught by a multi-tasking learning algorithm. Boccignone collected the joint parts of protrusion calculations in dynamic scenes by means of a mixture of Dirichlet Process as a basis for object-based visual attention. He also suggested an approach for parting a video into vistas based on protrusion display of a video. A key benefit of Bayesian Models is their ability to learn from data and their ability to be unified with many factors in an essential maneuver. For instance Bayesian Models can be statistically used for natural scenes or other features, attracting the attention.

5. Decision Theory Models

Interpretation of decision theory says that cognitive systems are deduced to make decisions about learned environment conditions, optimum in a decision theory concept (for example the least possibility of error). Gao and Vasconcelos have argued that it is for detecting protruded features that are the best detection of an interesting class among other classes of the image. Afterwards they defined top-to-down attention as a classification with the least expectation error. Specifically, with a set of known features $F = \{F_1 \dots F_d\}$ of a location ℓ and a C-labeled class, $\mathrm{C}_\ell = 0$ is related to the drawn samples of surrounded locations and $\mathrm{C}_\ell = 1$ is related to the drawn samples of a smaller central location, focused in ℓ. The following judgment of the protrusion is carried out by measuring the mutual information, calculated with $I(F, C) = \sum_{i=1}^{d} I(F_i, C)$. They used DoG and Gabor Filters, measuring the protrusion of a point as the KL Divergence between the histogram of filter response in the point and the histogram of the filter response in the surrounded location. In some articles this framework is used for bottom-up protrusion by mixing it with a processing of the surrounded central image. Moreover, they unified mobile features (image flow) between consecutive image couples so that their model would be considered for dynamic stimulus. They chose a dynamic tissue model by means of Kolman Filter to receive mobile patterns in a dynamic vista. Hereby we show Bayesian calculations, a particular case of decision theory model. Protrusion calculations in entire theory's event depend on measuring the minimum possibility $P(C = 1|F = f_z)$ of the target. Applying Bayesian Law, we have:

$$P(C_l = 1|F_l = f_z) = \sigma\left(log\frac{P(F_l = f_z|C_l = 1)P(C_l = 1)}{P(F_l = f_z|C_l = 0)P(C_l = 0)}\right)$$

Where $\sigma(x) = (1 + e^{-x})^{-1}$ is the Sigmund Function. The amount of correction logarithm inside Sigmund can be written in details below:

$$-logP(F=f_z|C=0)+logP(F=f_z|C=1) \\ +\frac{P(C=1|L=l_z)}{P(C=0|L=l_z)},$$

Which is similar to equation 5, through the following hypotheses:

1) $P(F=f_z|C=0)=P(F=f_z)$ and 2) $P(C=0|L=l_z)=K$ for some K statics.

Hypothesis 1 says that in the absence of the target feature distribution is like feature distribution for a set of natural images since most natural images do not have target, this hypothesis is not really probable. Hypothesis 2 simply says that in the absence of the target feature distribution have equal possibilities all over the image locations. This hypothesis seems a very mild one. Mahadeuan and Vasconcelos presented an algorithm without any supervisor for spatial-temporal protrusion, based on biological mechanisms of cognitive categorization, based on movement, which was in effect a development of discrimination protrusion model. Protrusion combination of the surrounding center with tissues' dynamic power causes their model to be suitable for very dynamic backgrounds as well as moving cameras. In the article by Gu et al, an active mapping was first calculated by extracting main image features and discovering meaningful objects from the image. A compatible network filter is applied to this mapping to produce significant Regions of Interest (RIO), which is the location, in which the location, related to these peaks, is activated and is estimated by a repetition algorithm. Attention focus repeatedly moves on the RIOs, discovered by a decision theory mechanism. The generated combination of eye staring was determined from a cognitive benefit function, based on cognitive costs and award, whereas temporal distribution of different RIOs was estimated by memory learning and fading. Decision theory models were very successful when used for computer sight such as classification while they gave much attention in staring prediction.

6. Information Theory Models

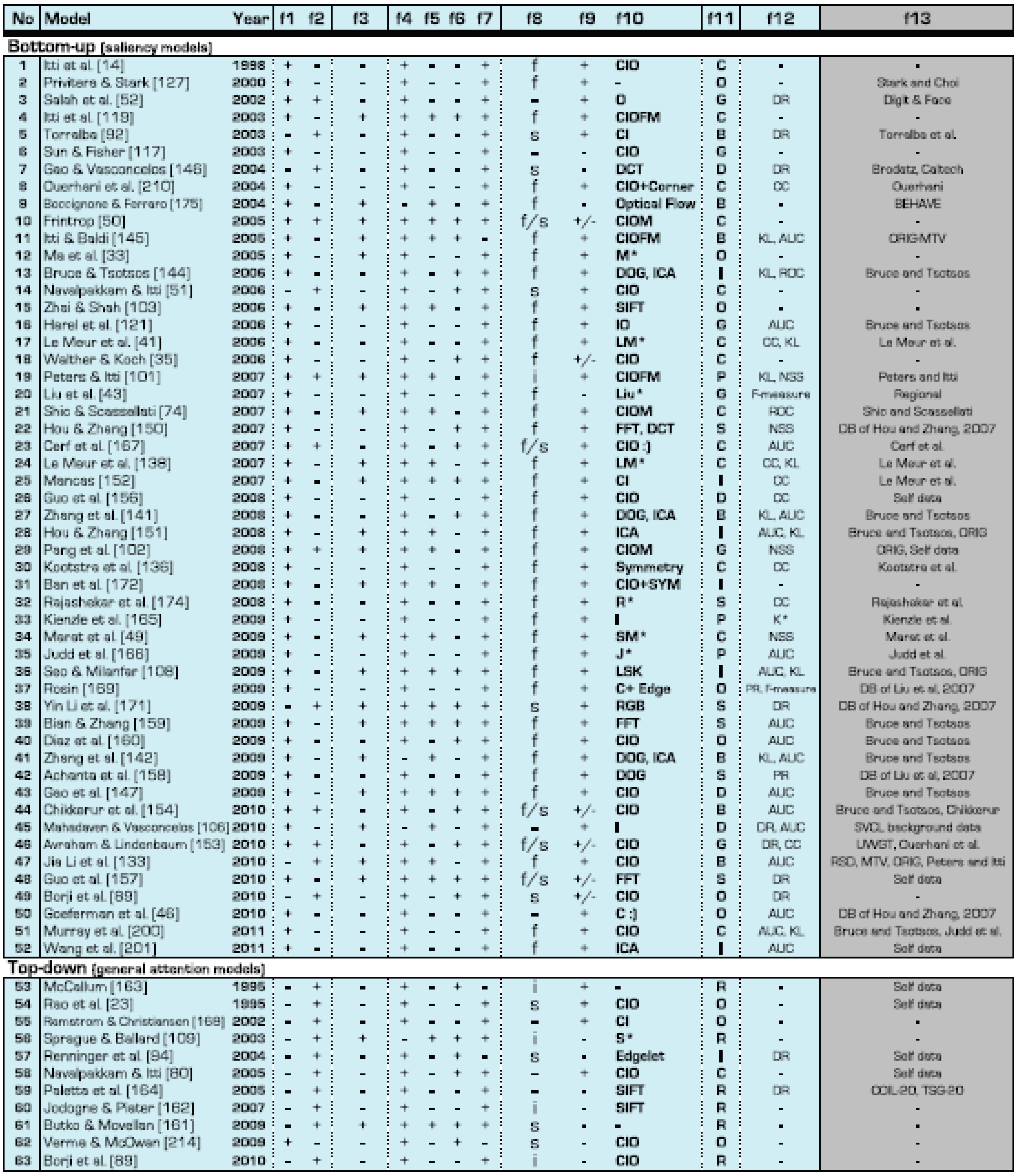
These models are based on a hypothesis, which focuses protrusion calculations in order to maximize the sampled information from some environments, dealing with selection of the important parts of a vista and omission of the rest. Rosenholtz has designed a visual search model, which can be used to estimate the protrusion on an image in free view. To do so, firstly the features of each p_i point is unified from a feature space (for instance uniform color space); afterwards the protrusion model defines the target as Mhalanobis Distance, Δ as the target's feature inter-vector, and T as the decomposing middle of the distribution, giving $\Delta^2 = (T-\mu)' \sum^{-1}(T-\mu)$. This model also exists below the noise measurement of a natural vista. Bruce and Tsotsos suggested AIM (Attention based on Information Maximization) Model which uses Shanon self-information criterion to calculate the protrusions of the image areas. An image's local area protrusion is information which other areas have concerning that surrounded space. Feature information of the image X is $I(X) = -\log p\ (X)$ which is attributed to the correction of X observations (P(X)) in reverse. To estimate I(X) the possibility function of the density P(X) should be estimated initially. In RGB images, considering a local piece in M*N, has bigger size than 3*M*N. To find ICA bases, they used a big set of RGB pieces, drawn from natural vistas. For an input image, the one-dimensional PDF has been calculated for each ICA basic vectors firstly by means of nonparametric density estimation. Afterwards in image areas the possibility of RGB amounts in a local piece of the image is produced from the related basic likelihood of ICA for that piece. Hou and Zhang introduced the ICL codifying approach of length increase in order to measure the disorganized benefit, related to each feature. The goal is to maximize the sample disorganization of visual features. By choosing features with high ICL, system calculation can obtain attention by choosing in both static and dynamic vistas. They have suggested ICL as a principle for the energy, distributed in attention system. In this principle, visual protruded signs are related to unexpected features. According to ICL definition, these features might extract disorganization benefit in cognitive state. Mancas has hypothesized that attention is attracted by small features in the image. The main operation of counting the sample image areas is done by analyzing the histograms, resulting in closer relation between this approach and Shanon's self-information criterion. Instead of comparing, only those pixels are separated which consider the spatial relations of the surrounding areas of each pixel (for example median and variance). Two kinds of the rare models are introduced: global and local. While global rarity considers the outstanding features on the entire image, some details of the image, resulting from local differences or rarity, might appear protruded. Similar to Center-Surround Idea (SDSR), initially a local structure of the image in each pixel is demonstrated by a matrix of local descriptors (local regression cores) which are resistant in the presence of noise and damage. Afterwards Cosine Similarity Maxim (a generalization of Cosine Similarity) is applied to the similarity of each pixel its surroundings. For each, the results of the protrusion mapping of its statistical likelihood shows the feature matrix Fi, the input of Fj Feature Matrixes of the surrounded pixels:

$$s_i = \frac{1}{\sum_{j=1}^{N} exp\left(\frac{-1+\rho(F_i,F_j)}{\sigma^2}\right)}$$

Here $p(F_\mathrm{i}, F_\mathrm{j})$ is Cosine Similarity Matrix between feature mappings Fi and Fj, and σ is a local weighted parameter. Columns of the local feature matrixes show the outputs of local command cores, modeled as below:

$$K(x_l - x_i) = \frac{\sqrt{det(C_i)}}{h^2} exp\left\{\frac{(x_l - x_i)^T C_l (x_l - x_i)}{-2h^2}\right\}$$

In which, $l = 1, \dots, P$ and P is the number of pixels in a local window and h is the softening parameters and C1 is Covariance Matrix, estimated from a set of spatial slope vectors, located inside the local analysis window around a sampling position.



No	Model	Year	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13
Bottom-up (saliency models)															
1	Itti et al. [14]	1998	+	-	-	+	-	-	+	f	+	CIO	C	-	-
2	Privitera & Stark [127]	2000	+	-	-	+	-	-	+	f	+	-	O	-	Stark and Choi
3	Salah et al. [52]	2002	+	+	-	+	-	-	+	-	+	O	G	DR	Digit & Face
4	Itti et al. [119]	2003	+	-	+	+	+	+	+	f	+	CIOFM	C	-	-
5	Torralba [92]	2003	-	+	-	+	-	-	+	s	+	CI	B	DR	Torralba et al.
6	Sun & Fisher [117]	2003	+	-	-	+	-	-	+	-	-	CIO	G	-	-
7	Gao & Vasconcelos [146]	2004	-	+	-	+	-	-	+	s	-	DCT	D	DR	Brodatz, Caltech
8	Ouerhani et al. [210]	2004	+	-	-	+	-	-	+	f	+	CIO+Corner	C	CC	Ouerhani
9	Boccignone & Ferraro [175]	2004	+	-	+	-	+	-	+	f	-	Optical Flow	B	-	BEHAVE
10	Frintrop [50]	2005	+	+	+	+	+	+	+	f/s	+/-	CIOM	C	-	-
11	Itti & Baldi [145]	2005	+	-	+	+	+	+	-	f	+	CIOFM	B	KL, AUC	ORIG-MTV
12	Ma et al. [33]	2005	+	-	+	+	-	-	+	f	+	M*	O	-	-
13	Bruce & Tsotsos [144]	2006	+	-	-	+	-	+	+	f	+	DOG, ICA	I	KL, ROC	Bruce and Tsotsos
14	Navalpakkam & Itti [51]	2006	-	+	-	+	-	+	+	s	+	CIO	C	-	-
15	Zhai & Shah [103]	2006	+	-	+	+	+	-	+	f	+	SIFT	O	-	-
16	Harel et al. [121]	2006	+	-	-	+	-	-	+	f	+	IO	G	AUC	Bruce and Tsotsos
17	Le Meur et al. [41]	2006	+	-	-	+	-	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
18	Walther & Koch [35]	2006	+	-	-	+	-	+	+	f	+/-	CIO	C	-	-
19	Peters & Itti [101]	2007	+	+	+	+	+	-	+	i	+	CIOFM	P	KL, NSS	Peters and Itti
20	Liu et al. [43]	2007	+	-	-	+	-	-	+	f	-	Liu*	G	F-measure	Regional
21	Shic & Scassellati [74]	2007	+	-	+	+	+	-	+	f	+	CIOM	C	ROC	Shic and Scassellati
22	Hou & Zhang [150]	2007	+	-	-	+	-	+	+	f	+	FFT, DCT	S	NSS	DB of Hou and Zhang, 2007
23	Cerf et al. [167]	2007	+	+	-	+	-	+	+	f/s	+	CIO :)	C	AUC	Cerf et al.
24	Le Meur et al. [138]	2007	+	-	+	+	+	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
25	Mancas [152]	2007	+	-	+	+	+	+	+	f	+	CI	I	CC	Le Meur et al.
26	Guo et al. [156]	2008	+	-	-	+	-	-	+	f	+	CIO	D	CC	Self data
27	Zhang et al. [141]	2008	+	-	-	+	-	+	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
28	Hou & Zhang [151]	2008	+	-	+	+	+	-	+	f	+	ICA	I	AUC, KL	Bruce and Tsotsos, ORIG
29	Pang et al. [102]	2008	+	+	+	+	+	-	+	f	+	CIOM	G	NSS	ORIG, Self data
30	Kootstra et al. [136]	2008	+	-	-	+	-	-	+	f	+	Symmetry	C	CC	Kootstra et al.
31	Ban et al. [172]	2008	+	-	+	+	+	-	+	f	+	CIO+SYM	I	-	-
32	Rajashekar et al. [174]	2008	+	-	-	+	-	-	+	f	+	R*	S	CC	Rajashekar et al.
33	Kienzle et al. [165]	2009	+	-	-	+	-	-	+	f	+	I	P	K*	Kienzle et al.
34	Marat et al. [49]	2009	+	-	+	+	+	-	+	f	+	SM*	C	NSS	Marat et al.
35	Judd et al. [166]	2009	+	-	-	+	-	-	+	f	+	J*	P	AUC	Judd et al.
36	Seo & Milanfar [108]	2009	+	-	+	+	+	+	+	f	+	LSK	I	AUC, KL	Bruce and Tsotsos, ORIG
37	Rosin [169]	2009	+	-	-	+	-	-	+	f	+	C+ Edge	O	PR, F-measure	DB of Liu et al, 2007
38	Yin Li et al. [171]	2009	-	+	+	+	+	+	+	s	+	RGB	S	DR	DB of Hou and Zhang, 2007
39	Bian & Zhang [159]	2009	+	-	+	+	+	+	+	f	+	FFT	S	AUC	Bruce and Tsotsos
40	Diaz et al. [160]	2009	+	-	-	+	-	+	+	f	+	CIO	O	AUC	Bruce and Tsotsos
41	Zhang et al. [142]	2009	+	-	+	-	+	-	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
42	Achanta et al. [158]	2009	+	-	-	+	-	-	+	f	+	DOG	S	PR	DB of Liu et al, 2007
43	Gao et al. [147]	2009	+	-	+	+	+	+	+	f	+	CIO	D	AUC	Bruce and Tsotsos
44	Chikkerur et al. [154]	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	B	AUC	Bruce and Tsotsos, Chikkerur
45	Mahadevan & Vasconcelos [106]	2010	+	-	+	-	+	-	+	-	+	I	D	DR, AUC	SVCL background data
46	Avraham & Lindenbaum [153]	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	G	DR, CC	UWGT, Ouerhani et al.
47	Jia Li et al. [133]	2010	-	+	+	+	+	-	+	f	+	CIO	B	AUC	RSD, MTV, ORIG, Peters and Itti
48	Guo et al. [157]	2010	+	-	+	+	+	+	+	f/s	+/-	FFT	S	DR	Self data
49	Borji et al. [89]	2010	-	+	-	+	-	+	+	s	+/-	CIO	O	DR	-
50	Goeferman et al. [46]	2010	+	-	-	+	-	-	+	-	+	C :)	O	AUC	DB of Hou and Zhang, 2007
51	Murray et al. [200]	2011	+	-	-	+	-	-	+	f	+	CIO	C	AUC, KL	Bruce and Tsotsos, Judd et al.
52	Wang et al. [201]	2011	+	-	-	+	-	-	+	f	+	ICA	I	AUC	Self data
Top-down (general attention models)															
53	McCallum [163]	1995	-	+	-	+	-	+	-	i	+	-	R	-	Self data
54	Rao et al. [23]	1995	-	+	-	+	-	-	+	s	+	CIO	O	-	Self data
55	Ramstrom & Christiansen [168]	2002	-	+	-	+	-	-	+	-	+	CI	O	-	-
56	Sprague & Ballard [109]	2003	-	+	+	-	+	+	+	i	-	S*	R	-	-
57	Renninger et al. [94]	2004	-	+	-	+	-	+	-	s	-	Edgelet	I	DR	Self data
58	Navalpakkam & Itti [80]	2005	-	+	-	+	-	+	+	–	+	CIO	C	-	Self data
59	Paletta et al. [164]	2005	-	+	-	+	-	-	+	-	-	SIFT	R	DR	COIL-20, TSG-20
60	Jodogne & Piater [162]	2007	-	+	-	+	-	-	+	i	-	SIFT	R	-	-
61	Butko & Movellan [161]	2009	-	+	+	+	+	+	+	s	-	-	R	-	-
62	Verma & McOwan [214]	2009	+	-	-	+	-	+	-	s	-	CIO	O	-	-
63	Borji et al. [89]	2010	-	+	-	+	-	-	+	i	-	CIO	R	-	-

Figure: 1-4 Evaluation factors of the models

7. Results

In this article we discussed the improvements of attention, focusing on protrusion models, which are performed recently, experimentally showing a set of past researches in a uniform way by qualitatively comparing the models. Improvement in this field greatly helps solving other challenging issues in terms of sight such as interpretation of scrambled and noisy vistas along with object detection. Additionally, there are many technical uses that can be done, as a result. Most previous researches of focus on bottom-up parts have visual attention. Although these former attempts are appreciable, visual attention field still lacks computational principles for task-based searches. A hopeful path for future researches is the development of models, taking work-based tasks into consideration, particularly in informal, complicated, and dynamic environments. Moreover, there is still no systematic calculations to understand hidden and obvious attentions, which must be clarified in future and there are issues beyond the field of computer sight that need the cooperation of machine's learning society.

Reference

[1] A. Borji, "*State-of-the-Art in Visual Attention Modeling*", IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 35 (2013), No. 1, January.

[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 (1998), no. 11, pp. 1254-1259, Nov.

[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, G.W. Cottrell, "*SUN: A Bayesian Framework for Saliency Using Natural Statistics*", J. Vision, vol. 8 (2008), no. 32, pp. 1-20.

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

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

[8] Wang, Y.-S., Tai, C.-L., Sorkine, O., and Lee, T.-Y., "*Optimized scale-and-stretch for image resizing*", ACM Trans. on Graphics (Proc. of SIGGRAPH ASIA) 27(5) (2008).

[9] Ma, M. and Guo, J. K., "*Automatic image cropping for mobile device with built-in camera*", in [IEEE Consumer Communications and Networking Conference], 710 – 711 (2004).

[10] Guo, Y., Liu, F., Shi, J., Zhou, Z.-H., Gleicher, M., "*Image retargeting using mesh parametrization*", IEEE Trans. on Multimedia 11(5), 856–867 (2009).

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

[12] SMR. Hashemi, M. Zangian, M. Shakeri, 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 (2015), No: 3 Special Issue.