Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 6 Issue 1, January - 2019

Automatic Analysis of Ocular Focus Detection Based on Visual Features

D. Oliveira Nascimento, Guilherme A. Oliveira, G. A. Wachs Lopes, Paulo S. Rodrigues*
Computer Science Department
Centro Universitário FEI
São Bernardo do Campo, Brazil
*corresponding author : psergio@fei.edu.br

***Abstract*— The human eye focusing is one of the most important tasks in the cognitive process of scene interpretation. The ability to estimate the focusing regions may vary according to the used algorithm and the image being analyzed, bringing a satisfactory efficiency in a specific set of images. This paper studies 9 methods proposed in the last decade, using 21 different features, discovering relations between the information within the images and the efficiency of the prediction. Using a supervised database, this paper shows that dispersion features for intensity data and color are more significant for the method efficiency than those based only on the average of the data. Besides, this paper proposes and analyses the capacity of Machine Learning techniques in identifying patterns inside the original images and selecting the most appropriate method to estimate focusing points.**

Keywords— Automatic Focus Detection; Focus Analysis; Computational Vision

I. INTRODUCTION

The large volume of digital information available has increased exponentially, mainly due to the progress of hardware and software technology. As the quantity of data grows, the request for a management for that data also grows.

Many of these digital data are in multimedia formats, such as videos, audios, and images. The presence of multimedia in our lives also grows each day, so does the need of automating visual tasks. However, these tasks are not so trivial to implement since the interpretation of the visual content is one of the major challenges.

At first, this interpretation takes into consideration only a few sets of low-level features contained in the image: color, intensity, shape or texture. Models based on these features showed a limited accuracy when compared to a human interpretation.

Although, it is known, based on neuroscience researches, that many other features are essential for a scene interpretation, some of them are: context, previous experiences, previous attention and so forth. All these new features require a deep study for a clear comprehension so they can be used on a computational model in an efficient manner.

Among the features that can be studied, there are human ocular focus regions, of which, in most cases, match with regions of interest [1. 2. 3. 4]. Because of that, for 25 years, computer scientists develop techniques to predict such regions.

The efficiency of today's automation techniques to predict the focal point vary according to the image class. However, it is still uncertain which image features causes the computational methods to have poor performance. Today it is possible to select the best method only when comparing with the human pattern obtained by eye tracking equipment.

Surprisingly, the focal mechanism that controls the eye movement is independent from the image processing in the temporal and parietal areas of the brain. Showing that, even blindly, the eye still focuses into determined areas, possibly attracted by features as color contrast or intensity. However, the temporal, occipital and parietal cognitive areas significantly influences the fixation, it is then a fundamental element for automated methods.

In the last decade, several methods appeared with the propose to predict the human focus point in an automated way, mainly in a natural scene. Although, it is still not clear which method is the best to predict those points.

Looking forward to a comparison between the methods, in this paper it was selected 9 principal methods chosen by your acceptance in the scientific literature. They were tested for 21 different features, under the same image database, making possible this comparison. It was made an election of the best method, accordingly to the class of the image it was submitted to.

In this paper will be presented four main contributions: first, the study of which image features influence the most in the efficiency of the prediction; the second contribution is the study of which are the present patterns in regions focused by humans; the third is the analysis of today's methods performance as they process complex images; the fourth contribution is the creation of a neural network to select the best method to be used.

II. Related Works

Researches on the Visual Attention areas have been evolving with the years and now it is now an active part in Computer Vision area. These researches try to reproduce automatically the human attention model. The term Attention correspond to all factors that influence the selection mechanism [5], visually refers to the areas focused by the human retina. With the advanced studies of the psychophysical area, it is known that the attention regions are related with the Regions of Interest (ROI), as presented in [1, 4, 6], showing the importance of the focus on reaching an objective, either it being an interpretation or the search for something.

Once the attention points are predicted, several applications appear, like: image segmentation [7], rendering [8], compression [9, 8, 7] and image testing for publicity [10], so forth.

However, even with the evolution of the studies and automatic methods of prediction of the focal points, some factors still make an efficient prediction difficult.

The study presented in [11] identified the existence of image classes of which the prediction techniques present a low consistency of the human visual focus. That is, they do not show a pattern among the observers. In both cases, the factors that cause this variation are unknown.

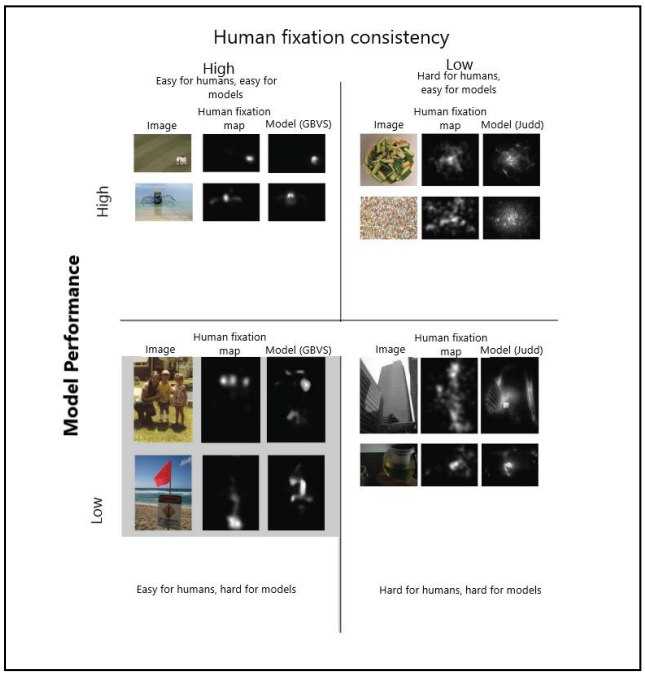

Fig. 1. Adapted from [11]

The Fig. 1 shows some images grouped by the performance of the methods of prediction and the human focal consistency. In the upper-left quadrant there are images where a big consistency on the focal points is found, but the automatic models presents a higher accuracy in the prediction. In the bottom-left quadrant humans present a good consistency, but the automatic models show a poor performance. In the upper-right, humans present a higher focus dispersion, like the automatic models. And in the bottom-right quadrant both present low consistency and performance.

The following discuss with more details the concepts and the related works of those prediction techniques.

There are 3 main mechanisms or attention models that influence the areas of an image that must be focused: bottom-up, based on features of the scene; top-down, determined by cognitive phenomena as the knowledge, previous experiences, expectations or objectives; and the hybrid models, using both bottom-up and top-down [5].

The bottom-up model is based on local features of the scene. For instance: color, intensity, orientation, symmetry, and others. The use of low-level features makes the bottom-up a fast and involuntary mechanism.

In the paper of Nothdurft [12], an example of a bottom-up attention is illustrated: an image of a horizontal bar among vertically oriented bars was presented to volunteers. It could be observed that the attention was immediately drawn to the horizontal bar. In this example, only local stimuli factors determined the human focus, free of the cognitive influence.

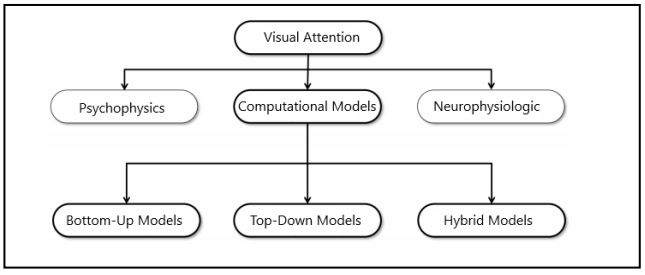

Fig. 2. Shows the hierarchy of the Visual Attention Models. Adapted from [5]

Although, there are other factors that influence a region to be focused besides stimulus. Those are cognitive factors, that are based on the context or the specific human task. The top-down model considers these factors in its prediction. Because of your complexity, the top-down model is slow [13, 14].

Examples of top-down elements that draw human attention are [11]: faces, parts of a face, human beings, texts, horizons, interaction between the objects and surprise elements. The focused region can also be influenced by a specific objective. A good example of the focal behavior oriented by an objective was the one identified in an experiment of Yarbus et al in [15], which presented to several observers an image containing people in a room. Each volunteer had been questioned previously about the features of the image that was going to be shown, for example: “What is the average age of the persons in the picture?”, “What is the average purchasing power of those people?” and others without any question. As a result, each group

presented a distinct eye movement. Those asked about the purchasing power focused on the wearing, those asked about the ages focused on the faces.

As both factors bottom-up and top-down influence the focused region, the latest prediction techniques use the hybrid model. They consider both cognitive and stimuli factors.

Next, the main prediction techniques described previously are presented with more detail.

The most common way to predict the human attention points is with Saliency Maps, which are images of intensity, where the regions with higher value (normally the brighter ones) have high probability of focus. The Fig. 3 exemplify original images with the overlapped fixation points and the saliency maps automatically generated respectively in the second row.

Fig. 3. In the first row examples of images with human fixation points. Adapted from [16]

The GBVS (Graph-Based Visual Saliency) is a bottom-up technique that uses the graph structure, despite the chain principle by Markov to predict attention points. This technique was proposed by Harel [17], and presents 3 main stages: (i) extraction, where is extracted a vector of features or saliency maps; (ii) activation, where is created a “activation map” using the feature vectors; and (iii) normalization, normalizes the activation maps and combine them in a unique map.

In the results of [17] this method presented an improvement in the capacity to predict the focal points.

The saliency model Context Aware try to detect regions of the image that are enough to interpret it. According to Goferman [18], the saliency regions can contain not only objects, but the background as well, with the condition that the background helps the identification of the context.

The Fig. 4 presents the difference between the concepts of object and saliency context. For the original images presented in the first line, users described the scene in distinct ways which are presented in the second line of the Fig.. Later, it was extracted from the image saliency objects (row 3) and the region showed by the Context Aware (row 4). Note that the background was used to interpret the image, except the first image.

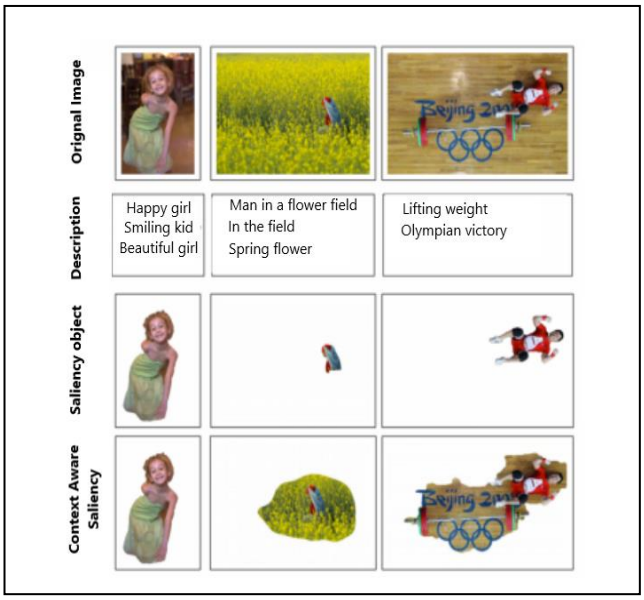

Fig. 4. Difference between objects and saliency regions to scene interpretation. In the first row are the original images. In the second, descriptions made by users. In the third row, the saliency objects. And in the last row the region given by the Context Aware method with context information. Adapted from [11]

To create a map that uses the background, the Context Aware is based on 4 psychology principles: (i) low level information, as contrast and color; (ii) global considerations; (iii) gravity center, in which the salient pixels must be grouped.

The Weighted Maximum Phase Alignment Model (WMAP) was proposed by [19] and is based on processes in frequency.

The saliency maps generated by the model are result of a sequence of processes realized in 3 stages: initial stage, analysis stage and system output.

In the first stage, the colors in the RGB system are decomposed in 3 less related channels, with the PCA technique. Later the Spectral Whitening filter is used in the frequency domain. In the second stage, it is analyzed the maximum alignment of the local level of each pixel. In the last stage, the three channels are combined in a unique saliency, and a Gaussian filter finalizes the map.

The model for creation of saliency maps Adaptive Whitening Saliency (AWS), proposed by [20], based on invariant features of scale and orientation.

In the saliency map creation progress, the channels Red, Green and Blue from the RGB system of the original image are split and processed individually. Later, each decomposed channel (Whitening transformation) is converted to the frequency domain, where its invariant features of scale and orientation are extracted. Returning to the spatial domain, through the inverse Fourier transform, the maps are joined and normalized in a single saliency map.

The central saliency map is unique to all images of the same dimension, not depending on any feature of the scene content for it creation. This map is computed assuming the maximum of 1 to the saliency value in the center of the image, and as the point distances from the center, its intensity decreases following the Gaussian equation below:

$$S(d) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(d)^2}{2\sigma^2}} \quad (1)$$

Where d stands for the distance, in pixels, from the analyzed point to the center of the image and σ is given by $\min(\frac{Height}{4}, \frac{Width}{4})$.

The central saliency maps can be applied as bias to the traditional prediction maps, combining them in a single map. This union occur with the following equation:

$$Map_{new} = \omega * Map_{central} + (1 - \omega) * Map_{saliency} \quad (2)$$

Where is an input parameter that represents the weight given to the central map after the combination.

In the study [11], it was identified that the central model raises the performance of most maps. Reaching the maximum point at ω = 0.8. This proves the high influence of the central map when focusing an image.

There are many features that can be extracted from images, some of them are: color, texture and intensity. It is possible to extract the same feature in different ways. Many attention methods created their own feature extraction model to predict focal points, such as not only those cited previously but also orientation [21]; symmetry [22]; texture [4]; face detection [23]; entropy [24]; Wavelets [25]. Hence each method is specialized in predicting attention points from different image classes.

Intuitively, humans need distinct features to recognize Regions of Interest, according to the image. In the Fig. 5 are shown some examples of different images. Feature extraction can also be used to comprehend and group classes of images. In this paper some features are be used in a learning process to discriminate and classify images in groups accordingly to the method prediction accuracy on them. Next, the image features used in this paper will be described.

Intensity is one of the most used features in the image processing and correspond to the gray scale of an image. The most traditional way to represent it is through the histogram of grayscale intensities. One extension of this feature is the color space, frequently modeled as a 3-dimensional space as RGB and HSV.

Another common metric is the area of a region and can be computed counting the number of pixels inside a certain area.

One can use the color/grayscale histogram as a probability distribution and compute the entropy as a metric to evaluate how distributed the information of the image is. The Shannon entropy on a gray scale image is given by $S = -\sum p(X)\log p(X)$, where X is the array of region intensity and p(X) the probability to occur each intensity value.

Fig. 5. Diversity of images and features.

In a binarized image taken from a saliency map one can count the number of disconnected regions to find how many regions are in focus.

All these described methods play an important role to describe information from image. However, these metrics cannot be taken alone without some learning algorithm. In the next paragraphs, it will be described some learning techniques. Machine Learning is a field of Artificial Intelligence that is composed by several algorithms and techniques that given a certain data can learn, or, improve the execution of a task or problem solution. These algorithms are classified in some distinct models, defined by the learning process characteristics. The main models are: (i) supervised learning, (ii) unsupervised learning and (iii) reinforcement learning. [26]

In Visual Attention, few applications using learning techniques were used to predict focal points. It was used in the paper proposed by Brecht e Saiki [27], where the Itti and Koch [28] model were extended implementing neural network.

Another possible application of machine learning is to evaluate the most accurate technique to create a saliency map of a specific image.

One of major advantage of a quantitative appraising methods in image digital processing is that one can eliminate subjectivity and make and standard efficiency criteria, which makes possible the comparison between methods in different works.

In the MIT platform Saliency Benchmark, are shown several comparisons between the methods under 7 different appraising methods. Next are presented the most used methods: AUC, CC, And NSS.

The Receiver Operating Characteristic (ROC) is a graphical representation of a binary classification performance [29]. The behavior of the ROC curve indicates the accuracy of the classifier as its sensibility is changed. The AUC (Area Under Curve) is commonly used as an appraiser of the system or the tested model. Assuming 1 to the maximum efficiency and 0.5 to random classification.

In the saliency map, each pixel over a threshold is considered a point where the ocular focus occurs, otherwise there is no focusing. Correspondingly in the

human map, the pixels of the same location are considered as a base to classify the point as a focus point (positive) or not (negative) [30. 23].

Correlation Coefficient (CC) is a common metric used to compare two images. When the CC between a saliency map S and the golden pattern G is computed, it is possible to find how correlated these two maps are. The CC between the maps is given by:

$$CC(G,S) = \frac{\sum_{x,y}(G(x,y) - \mu G) - (S(x,y) - \mu S)}{\sqrt{(\sigma^2 G * \sigma^2 S)}} \quad (3)$$

Where μG and μS are the average of every pixel in the human map G and the calculated S, respectively. $\sigma^2 G$ and $\sigma^2 S$, respectively, the variance of the G and the S.

One characteristic of CC metric is that it ranges between -1 and 1, respectively a complete inverted correlation or a direct correlation. When no correlation is found, CC is 0.

The NSS (Normalized Scanpath Saliency) [31] evaluate how distinct the focus region is in relation to the whole map, allowing the inference to the saliency map accuracy. Saliency maps that have high intensity in both focus region and in others image areas do not have a satisfying NSS. It indicates that all regions are being focused, and that is not true. Ideally, a good saliency map has high intensity in the focus regions and low intensity in the rest of the regions. The NSS can evaluate the precision of a saliency map.

The Fig. 6 shows a good example of NSS. The first stage to be executed is the transformation of the saliency map S into a normalized S_{norm} with average of 0 and standard deviation of a unit, following the equation below:

$$S_{norm}(x,y) = \frac{S(x.\ y) - \mu S}{\sigma S} \quad (4)$$

Where the σS is the standard deviation of S.

With the normalized map, the average of intensities in the focus points of the golden pattern is extracted. Given that F is the set of n points focused by humans, the NSS value is given by:

$$NSS = \frac{\sum_{i=1}^{n} S_{norm}(F_i)}{\sigma S} \quad (5)$$

If the NSS value is high, the map precision is also high. If the value is equal to 0, the prediction level will be random, because it is in the average of the others image areas.

The NSS value represents how many standard deviations the average of the focal points is above or below the random level. Thus, there are no direct relation between the NSS metric and a quality value. This measurement should be used comparatively between methods, indicating the most accurate of them.

III. Proposed Methodology

In computer vision area several methodologies were developed in order to automatically identify which areas will be focused by humans during a scene interpretation. Although, none of them stands out from all image classes, since the prediction accuracy vary according to the technique and the image type [11]. A methodology that support the factors that influence the focus and the selection of the most appropriate prediction model according to an image, improves the prediction capacity of the focus regions.

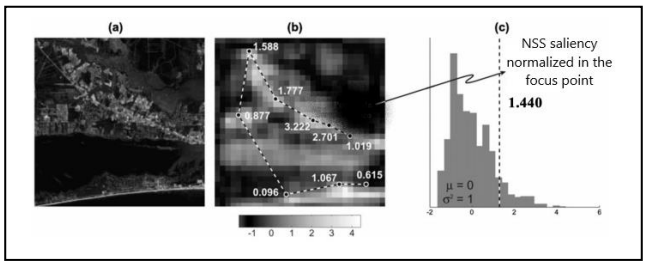

Fig. 6. An illustration of NSS method. (a) Original image. (b) Normalized Saliency Map at an average of 0 and standard deviation of 1. (c) Intensity histogram of the Saliency map and the average on the focusing points. In the example, the points that correspond to a human focus are 1.440 standard deviation above the average. Adapted image from [31]

Fig. 7 presents a methodology proposed in this paper. This methodology creates information for 2 main tasks: (i) features analyses thought the search of features patterns within the scene; and (ii) method selection, which is based on a neural network classifier trained to find out what is the best technique to analyze a scene. With this architecture, it is expected to obtain the information needed to comprehend how the features within the scene influence the attention model and the computational prediction techniques.

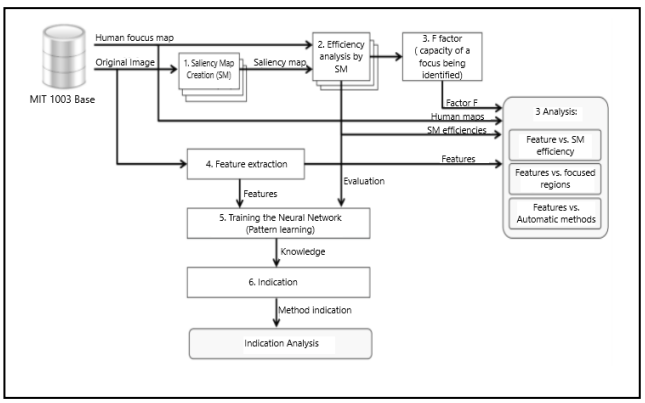

Fig. 7. Methodology architecture.

Moreover, we expect to increase the efficiency of the predictions through the automated selection of the best model to generate saliency maps.

Next, we will describe the database used in this work and the stages of the methodology.

In this paper, it was chosen a supervised image database which enables the comparison of the results of this paper with others of this area, also allowing a quantitative evaluation when compared to the human pattern.

Therefore, it will be used the images from the Massachusetts Institute of Technology (MIT) database [32] composed by 1003 images in which 1000 of them are natural images and 3 are artificial. All images have a respective human saliency map with data of the attention points, captured through the monitoring of the observer's eye during the interpretation.

In Fig. 8 some examples of images of this database are shown with its respective saliency map. Nowadays, there are others image databases with focal points capture, as presented in [22, 33]. However, they have a reduced number of images or have a specific context, so it cannot be used to have a more general training.

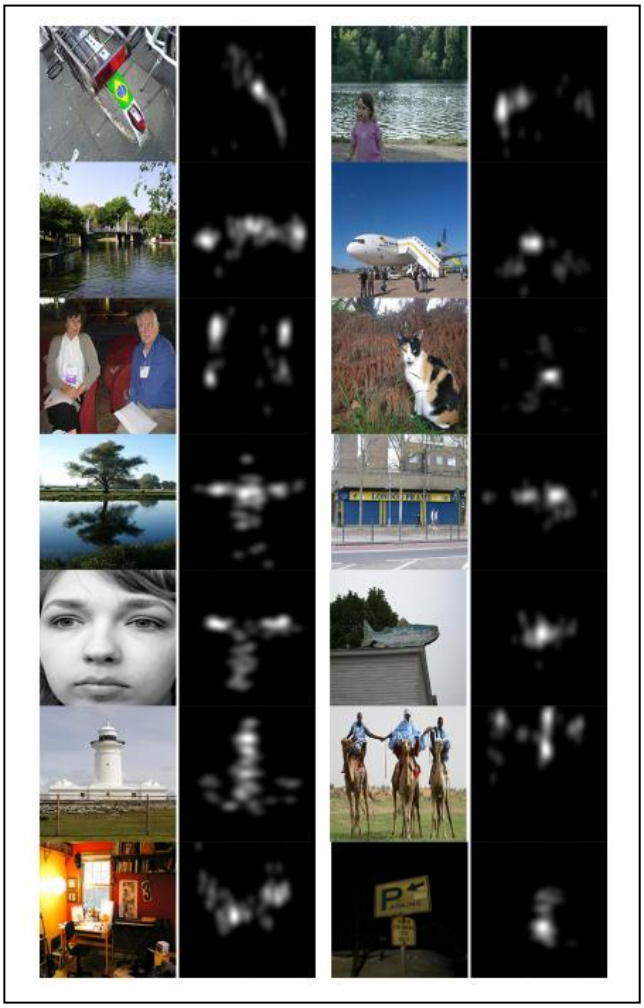

Fig. 8. Sample of the images from the database used in the paper and its saliency maps.

The goal of the first stage of the methodology is to create saliency maps for each image of the database. Altogether, there will be 9 saliency maps for each image, one obtained from each method.

Therefore, for each image I_k of the training set of images I = $\{I_1, I_2, \dots, I_n\}$, 9 saliency maps will be generated $S_{i,k}$ (where i ranges from 1 to 9 and k ranges from 1 to n), 5 automatic saliency maps (AWS, Context Aware, WMAP, Graph-Based Visual Saliency (GBVS) and central map) and 4 combinations of the same 5 maps with the centralization map, except itself. It will be used weight w = 0.8 for the centrality map in the combination with the others.

All maps will be normalized after its creation so they have intensity in the interval [0, 1]. Therefore, the most intense point in the map will have value of 1 in the normalized map, and the less intense point will have value of 0.

The 9 saliency maps generated in this stage intend to point the regions that was focused by humans. However, to confirm its accuracy, these maps will have its efficiency analyzed using the criterion described in the next paragraphs.

The prediction accuracy of an automatic focal points map varies with the used technique and the analyzed image. To quantify this accuracy, it is necessary to compare each output with the pattern generated by humans, captured with an eye tracking equipment.

In this stage, the maps generated in the previous stage will be evaluated on how much they were close from the human pattern. So, the saliency maps $S_{i,k}$ will be compared with the golden pattern G_k regarding image I_k. Each map $S_{i,k}$ will be evaluated by 3 distinct efficiency metrics: AUC, CC and NSS.

In order to put all these metrics into a single one, we purpose Equation (6) as a final efficiency measurement. In this equation, the values of AUC are normalized by the expression |AUC * 2 - 1|, once the AUC ranges between 0 and 1, and 0.5 represents total randomness of the system. Thus, the original values as 1 (directly related) remains as 1, and 0 turns into -1 (inversely related).

Another measure that might be normalized is the CC, once its values are between CC=1 (directly related) and -1 (inversely related), and 0 when no relation is found. The normalization assumes the value of |CC|.

On the other hand, the NSS can assume values in the real domain. Therefore, it is used as a weighting factor of efficiency in a high precision case of the saliency map.

The efficiency function $Ef(I_j, M_k)$ of an image I_j and a method M_k is given by the average of the normalized values of AUC and CC, weighted by the NSS measure.

$$Ef(I_j, M_k) = \frac{|AUC * 2 - 1| + |CC|}{2} * NSS \qquad (6)$$

The measures AUC, CC and NSS are calculated between the two maps, $S_{i,k}$and G_k. Since the accuracy of all methods with all images of the database will be obtained, it is possible to conduct a supervised learning algorithm predict the technique that gives the best focal points for a specific image.

Besides the learning algorithm, we can create a unique factor of prediction.

That is, as previously said, it is known that there is a set of images of which the computational models have, generally, low prediction performance [11], even if each method has its own independent efficiencies $Ef(I_j, M_k)$.

In order to identify the general efficiency of an image, the F factor was created. In this paper, the image factor of an image I_j is given by $F(I_j)$, and is expressed by the average of all $Ef(I_j, M_k)$:

$$F_{I,j} = \frac{\sum_{k=1}^{m} Ef(I_j, M_k)}{m} \qquad (7)$$

where m is the total of models present in the set of all methods M.

So, the factor F represents the capacity of an image to have your focal points automatically identified, and without the use of the golden pattern, by computational attention models. F can assume values from the real numbers and must be analyzed with other images.

Several features can be used to describe an image, from features related to low level information to high level information. In feature extraction stage of the methodology, low level features will be extracted from the images of the MIT1003 database in order to represent and analyze them.

Altogether 21 features will be used on this paper:

1. Intensity - given by the average $(\bar{I})$ and standard deviation (σ) of a gray scale image intensity level;

2. Color - for each channel on the HSV, the following averages will be analyzed (CH, CS and CV) and their standard deviation (σCH, σCS σCV);

3. Entropy S - calculated over the intensity histogram;

4. Number of Regions - number of regions disconnected over a threshold θ;

5. Area - percentage of the area size over a threshold θ in relation to the image size.

The θ values used on the number of regions and area features are given in two different ways: (i) Adaptive - determined by the division of the intensity histogram by the K-means method (K = 2); (ii) Predefined - independent of the image, a threshold occur in the values {0.97, 0.95, 0.90, 0.70}, as it was used on Judd's paper [34].

Then, measures related to the number of regions and the area size are appended:

1. Growth rate (regions) - given by the ratio of the quantity of existing regions on the image when limited to the thresholds of 0.70 and 0.97, respectively.

2. Growth rate (area) - given by the ratio of the size of existing regions on the image when limited to the thresholds of 0.70 and 0.97, respectively.

Overall, these 21 features will be analyzed on the experiments and so the input for the training network.

It is known that some techniques are better in performance over a determined image class. Although, to determinate which technique is the indicated to a certain image without comparing to the human pattern is a difficult task. The training of the neural network creates a neural network with multiple intermediate layers to learn which method is the most indicated to predict the focal points of and specific image.

With this in mind, 4 stages will be executed: the creation of the saliency maps; performance analysis; feature extraction; and the pattern training.

The Fig. 9 shows the topology of the proposed network. The input layer will be composed by a set E of 21 neurons, where each neuron E_c correspond to an input of the feature C extracted from the original image. The output layer will have 9 neurons referring to the 9 studied methods: AWS, Context Aware, WMAP, Graph-Based Visual Saliency (GBVS), central map, AWS + central map, Context Aware + central map, WMAP + central map and GBVS + central map.

During the learning, the output neurons will have their errors computed based on the efficiency (E_f), calculated previously by the evaluation step. Setting the best method with 1 and the others with 0.

The definition of the network topology many times occur in an empiric way and is adjusted manually. However, on this paper we will create several random topologies inside a specific interval that do not exceed an execution time. There will be topologies with 3 layers and 50 neurons by layer. The network with less error will be preserved for the next steps of the methodology.

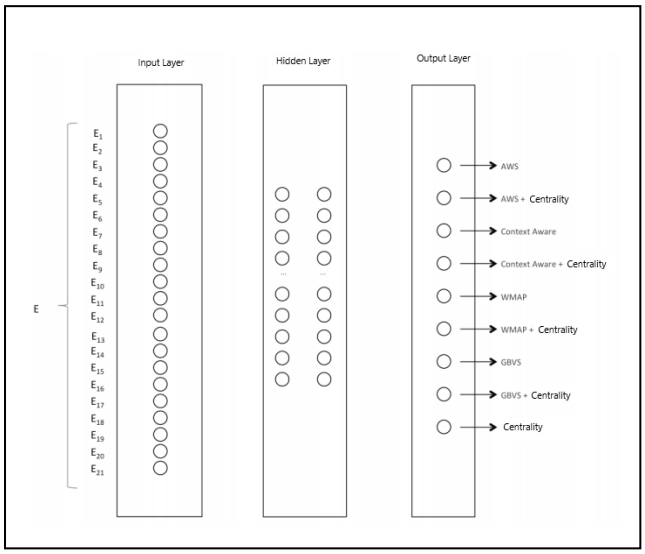

Fig. 9. Representation of the neural network training.

Once the training step build up the network, it can be used to suggest the best prediction method for any image. Note that, once the network is trained, even images that are not in the trained database can be used as input.

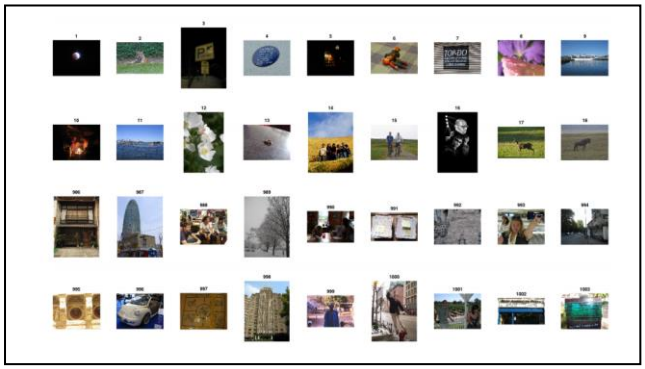

Fig. 10. Images with high predictability factors (From 1 to 18) and images with low predictability factors (from 986 to 1003).

Next, the indication stage is used to select the method that tends to have the best SM (Saliency Map) to a determined image, without the help of the human golden pattern. Two main steps occur in this stage: feature extraction and pattern recognition.

All the features will be taken in the same way that occurs in the Training module. Therefore, it is possible to insert data on a previously trained neural network and compute the values of the output layer, using the feed forward process. Each neuron of the output layer represents the saliency map of an automatic method. Thus, it is possible to assume that the most indicated method is the one that has the higher value on the output neuron.

After all the stages of the proposed methodology, it is intended to achieve the information necessary to understand how the 21 features influence the attention model and the computational prediction techniques. This also make possible to increase the prediction efficiency through the automatic selection of the best model to generate the saliency maps.

IV. Results And Discussion

On the creation of the SM, 9 distinct prediction maps were generated to each image on the MIT1003 database. These SM are compared to the human pattern and evaluated on how close they are from each other.

The first two rows of Fig. 10 illustrate examples of images with highest predictability (F) (Fig. 10 from 1 to 18). In the same way, the last two rows illustrate images with the smallest factors (Fig. 10 from 986 to 1003).

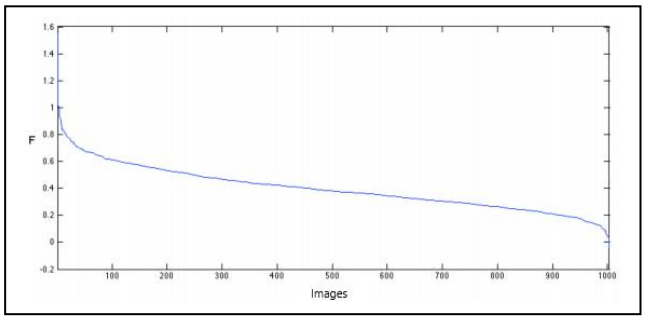

Fig. 11. The sorted images of MIT base with the F factor.

For each image of the database, it is computed the prediction efficiency factor (F) as explained in Section 3. Keep in mind that F factor represents how much an image can be better predicted in relation to the focus given by the prediction methods, the greater the F factor is, the easier the image can be predicted. Similarly, the smaller the F factor is, the harder the image can be predicted.

The Fig. 11 shows the curve of the F factor on images from the database, that is, on the horizontal axis the images were sorted decreasingly in relation to F, which has its values on the vertical axis. According to this Fig., it can be observed that the decay is not too intense. Moreover, there is an initial peak and a final peak. This suggest that there are images that its focal region is really easy to predict, while in others the focal region is really hard. As an example, the Image 1 on the Fig. 10 has its background and the object well defined. On the other hand, the Image 1003 does not seem to have a homogeneity of its objects in the scene.

Considering all the 21 features, it can be observed how they behave on each image of the database. With this in mind, each one of the 21 subplots of the

Fig. 14 show the behavior of each feature individually, as the F factor decreases.

Each of these subplots in Fig. 14 was built keeping the same order from the Fig. 11. However, on the vertical axis it can be observed the value of the respective feature.

The features can have different behaviors for each image. For instance, Fig. 14 (a) shows the values of the average intensity feature, which is an example of low correlation with F. However, the Fig. 14 (f), refer to the standard deviation of the color values on the H channel of the system color HSV, which is an example of high correlation to F. Another example is the Fig. 14 (i), corresponding to the entropy on the image intensity histogram. Both of them show some relation with the factor F; that is, once each subplot has on the horizontal axis a decreasingly curve, the studied features tend to relate positively with the factor F. The opposite is also true. So, the standard deviation of the values on the channel H and the quantity of regions over the threshold of 0.95 are inversely correlated to F.

Other examples of features with high correlation in the Fig. 14 are: (b) intensity deviation, (h) channel values deviation V, (j) quantity of regions defined by an automatic threshold. The relation between the features studied and the factor F can be defined quantitatively by the correlation coefficient (CC). The CC can have values between -1 and 1. Assuming the value of 1 when the studied variables are totally related. A correlation totally inverse receives the value of -1. However, when the CC reaches the value of 0 no relation exists.

It is possible to analyze the CC in absolute values (|CC|), this shows how related the variables are, independently if it is a direct or inverse relation.

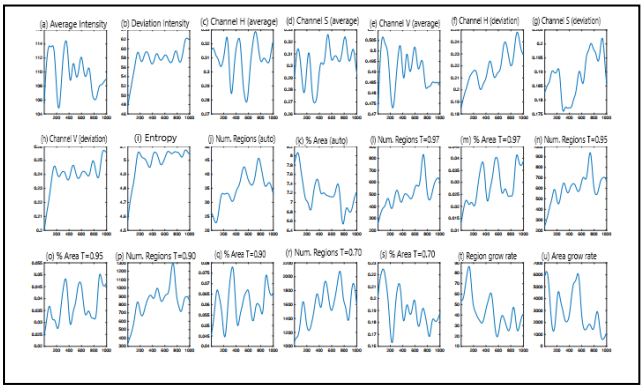

Fig. 12. Behavior of the studied features on the images sorted by its average.

The Table 1 shows all the features studied and the respective CC with the factor F, sorted decreasingly by |CC|. Observing the table 1, it can be observed that the feature that has more relation with the level of predictability of an image is the standard deviation on the channel H, but in an inverse manner. This suggests that the smaller the color diversification in an image, the easier is to automatically predict the human focus. The same way, the higher is the color dispersion on the channel H in an image, the bigger are the difficulties to predict the focal regions.

Besides the information on the color diversification, other dispersion features present as the most related with the level of predictability on an image, as the entropy of the intensity level and the standard deviation of the gray scale image and the channel V. All of them inversely related. On a subjective way, the high relation existing between these features happens because they represent the complexity of the scene with a lot of information diversity, despite of being used as base on the most of the automated methods.

Next to the most correlated features, there are three features related to the number of disconnected regions. The first is when an image is "binarized" to an automated threshold and, the other two are images are thresholded with the fixed values 0.95 and 0.97. This suggests that when there is a spatial distance between the region of interest in a scene, the harder it is to the human focus detection. The opposite is also true.

TABLE I. PRESENTS THE CC BETWEEN THE STUDIED FEATURES AND THE PREDICTABILITY FACTOR F. ROWS ARE SORTED DECREASINGLY BY THE ABSOLUTE VALUE OF |CC|.

FEATURES	***CC***	***|CC|***
H (Deviation)	-0.842	0.842
V (Deviation)	-0.776	0.776
Entropy	-0.762	0.762
Intensity (Deviation)	-0.760	0.760
Number of Regions (auto)	-0.748	0.748
Number of Regions (T=0.95)	-0.745	0.745
Number of Regions (T=0.97)	-0.735	0.735
% Area (auto)	0.706	0.706
Number of Regions (T=0.90)	-0.658	0.658
Number of Regions (T=0.70)	-0.658	0.658
Region growth rate	0.648	0.648
% Area (T=0.97)	-0.639	0.639
% Area (T=0.70)	0.626	0.626
Area growth rate	0.570	0.570
S (Deviation)	-0.508	0.508

FEATURES	*CC*	*|CC|*
% Area (T=0.95)	-0.507	0.507
Intensity (Average)	0.396	0.396
S (Average)	-0.352	0.352
V (Average)	0.312	0.312
% Area (T=0.90)	-0.300	0.300
H (Average)	0.006	0.006

Analyzing the opposite values of |CC|, the average of the H channel stands out. This shows the independence of the level of the color dispersion to the focal prediction. The same occur to the average of the intensity and color on the channels S and V.

Despite the relation with the factor F, the features can have relation between themselves. The Fig. 13 shows the CC between the studied features. It can be observed as highly related features the standard deviation V with standard deviation of intensities and entropy with standard deviation of intensities, respectively 0.980 and 0.872. Although, the relation with the standard deviation of H is on another level, 0.622, the same relation happens between the dispersion of the channels H and V. This suggests that color and regions with shadow of an image contribute on distinct forms to the focal prediction.

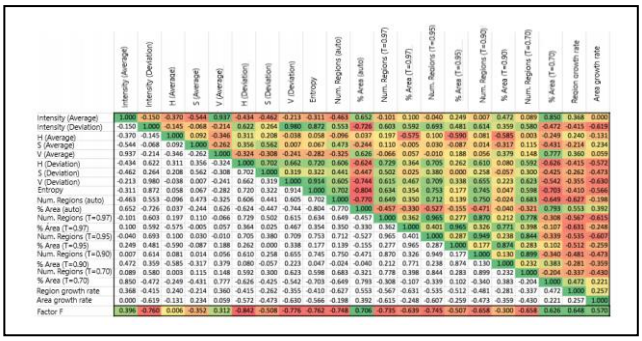

Fig. 13. CC between the studied features and the F factor.

It is also observed a highly relation between the fixed thresholds features and the number of regions. The same occur with percentage of areas and threshold changes. In an informal way, when analyzing the behavior of an image feature, it is possible to identify how they influenced the capacity of the method to predict the focus regions of the image. However, studying how these features behave on the focus points and in the rest of the scene can help us to comprehend what influences the focus to embed in automatic methods.

The Fig. 14 shows the comparison between 11 features inside and outside

the human ocular focus area, on the images sorted by the factor F. The graphic on the left side of the Fig. 14 (a) presents two overlapped curves. The green curve represents the behavior of the feature Average Intensity on the foreground, region focused by humans. On the other hand, the red curve represents the same feature on the rest of the image (background). The graphic immediately on the side (blue curve) represents the value given by the relation $C_f \oplus C_b$.

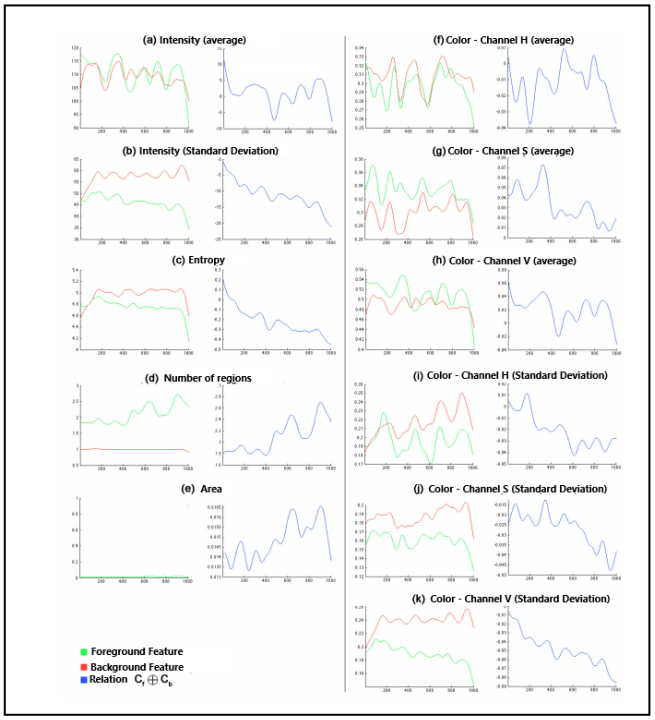

Fig. 14. Feature behavior studied on the focus region and outside it.

It is possible to observe on this Fig. that the features do not have significant differences in the behavior inside and outside the human focus regions, as presented in the Fig. 14 (a), Average Intensity. It is observed that the average in the focus region and in the rest of the image are close to each other and there is no pattern related to the F factor. The same occur to the value averages on the channels H (f) and V (h). However, there is a tendency that suggests that, for that database, humans tend to focus on regions where the color are brighter, as shown in the Fig. 14 (g). It is also possible to observe that as the saturation of the foreground and the saturation of the background get closer (blue curve), it is harder for the methods to generate the saliency maps.

Some other feature behavior are: humans focus in regions with low entropy Fig. 14 (c) and with lower intensity standard deviation Fig. 14 (b) and on the 3 channels of color HSV, respectively observed in Fig. 14 (i), (j) and (k). Although, as the difference of

deviation and entropy between the focus regions and outside raises, the less efficient the automatic methods of prediction are.

This suggests that other factors can be influencing the detection of attention points, such as spatial features. This can be observed in the Fig. 14 (d), where the number of disconnected regions focused by humans raises and the saliency maps get less efficient. This is related to the size of the focused area (e), once the human eye passes through a bigger image region to interpret it.

Knowing the behavior and the performance of each prediction method is as important as knowing how the features on the images or the focus region behave. Identifying which classes of images each method has better performance can improve the overall results.

The Fig. 15 shows examples of images and the methods that generate better saliency maps. Each image from the MIT database was grouped into columns accordingly to the method that generated the most efficient saliency map for it. Among the images from a certain group, it was sorted accordingly to the factor F. The biggest 5 efficiencies of each group were placed in the Fig. 15(rows).

Fig. 15. Example of images grouped by the most efficient saliency map (columns). On the rows, decreasingly, are presented the best 5 images in efficiency.

The Fig. 16 presents the feature behavior on the images from the same groups of Fig. 15. In each of these plots, the vertical axis corresponds to values of the features of each image in the respective group, sorted decreasingly by the factor F.

It is possible to verify on this Fig. that the Channel H presented a relative low value of standard deviation, between 0.20 and 0.24, for all methods. Furthermore, the images that the central method had better performance always have a high number of regions and high entropy. That suggests that this method is the best to images with more elements and regions.

The Fig. 17 shows a curve for each method. On the horizontal axis, the images are ordered decreasingly by the factor of focal predictability F, and in the vertical axis the values of efficiency Ef of each method, of which compares how close the saliency map generated by the prediction model is from the human map. Is possible to observe that the central model has the worst performance on images with higher F factors, although its decay is weaker than the others. Besides, on the harder images to predict the attention points, the best methods converge to the same level of the central model.

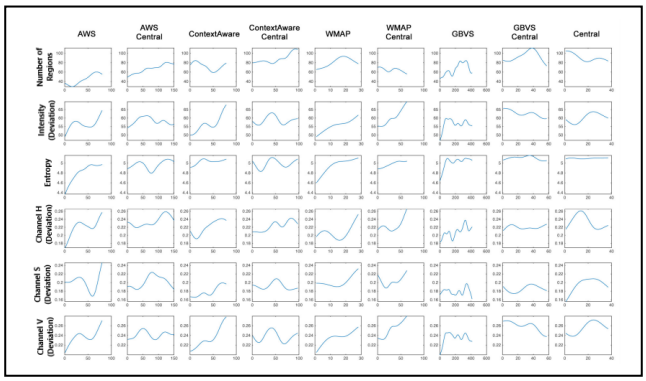

Fig. 16. Feature behavior (rows) among all images that had the same method (column) as the most efficient SM generator.

Such behavior indicates that, for complex images, the best a method can signalize is that the focus must happen in the center of the image, otherwise the effects of the model end up being less efficient. This behavior can be visualized on the Fig. 17, where all automate methods have a lower efficiency when compared to the central method. Until the inversion point, the studied automated models contribute to generate saliency maps closer to the human map, however, after this point, the efficiency tends to have the opposite effect: they are worse than simply indicate the center of the image as the region to be focused. Besides, there is a small difference between the automated methods with the center method, which indicates that the proposed models are not so reliable, the efficiency is due to the central model.

Furthermore, one can verify the point when the central model starts to make a positive difference on the saliency maps. The points represented by circles indicate the when the methods without the central model crosses the respective methods with the central model. When this top-down crossing of the method without the central model occurs, it shows the when the centrality feature tends to influence positively the prediction model. This behavior can be seen in all models, except in WMAP, where the centralization factor always tends to contribute positively.

With these results, it is suggested that the human focus is very influenced to the center and that this feature must be used to images with less efficiency.

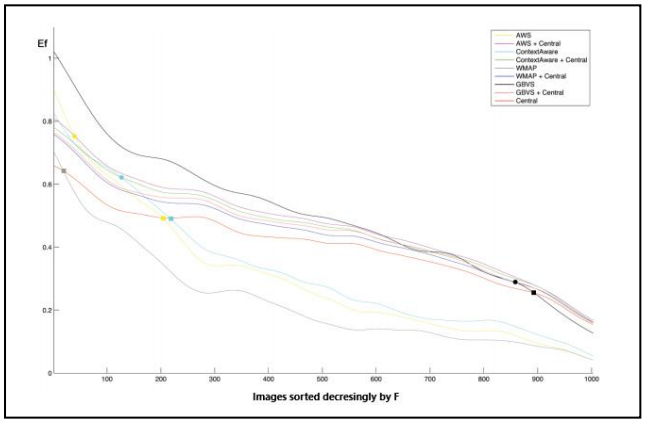

Fig. 17. Efficiency by method. In the horizontal axis images are sorted by its F factor, in the vertical axis the efficiency values Ef of each method.

Considering that the saliency maps have distinct efficiency depending on the image and the method used to generate it, knowing how to indicate the best prediction method to a determined image without the human pattern can raise the efficiency of the applications that need to predict the ocular focus.

With this in mind, in this paper it was proposed a neural network specialized to indicate which method is recommended to generate the saliency map. The topology of this network has 21 input neurons, 2 hidden layers, respectively with 40 and 20 neurons and activation function. Besides, the network has 9 neurons on the output layer, each one corresponding to a method. Thus, it is assumed that the most indicated method is the one with higher value on the output neuron.

The Fig. 18 presents the accuracy of the proposed network. The horizontal axis represents the ranking of the indicated saliency map, in relation with the other SM. The vertical axis presents the accumulative percentage of the number of indications of the network until the referent classification. For instance, 47.96\% of the methods indicated by the network were actually the best possible. In an accumulative way, it is observed that 56.63\% of the indications of the neural network were between the first and the second best method. The same way, 63.01\% of the network output are among the 3 best and so forth. Once more than half of the indications do not point to the method with the best efficiency, it is suggested that there is a difficulty for the network to indicate the most appropriate method.

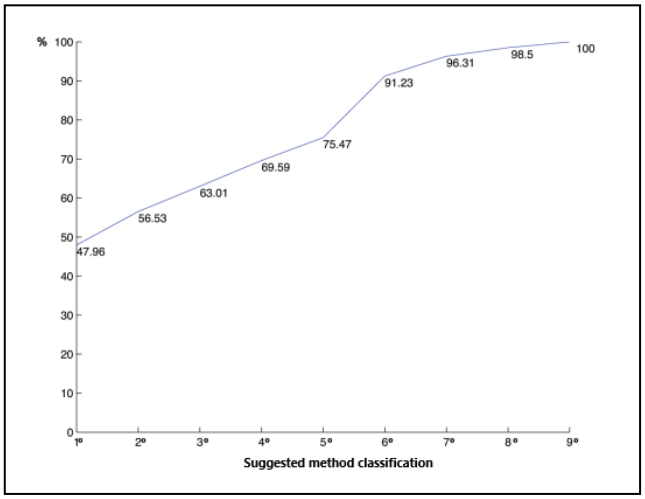

Fig. 18. Curve with the accumulative percent of the indication ranking.

The Fig. 19 shows the histograms of images by the focal prediction method. In Fig. 19 (a) there is the distribution of the images on the methods that generated the most efficient SM. Notice that the model GVBS is the best in performance to most of the MIT database, followed by AWS and Context Aware, both with the central model. On a similar way, it is expected that the indicated methods followed a similar distribution. Although, in Fig. 19 (b) it is verified that most of the output from the network indicates the GBVS method. This suggests that with the input features on the network, it was not possible to establish a separation hyperplane. So, the network chose to select the same method almost always, trying to minimize the error on the training stage.

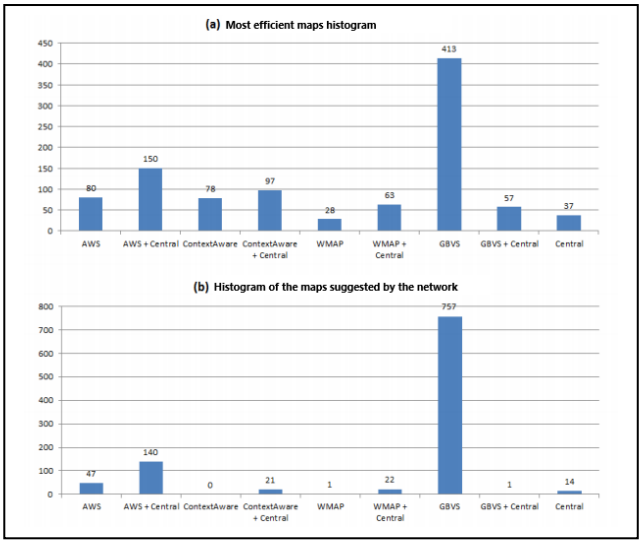

Fig. 19. Histogram of (a) the most efficient SM when compared to the human pattern and (b) the network response histogram.

Thus, this behavior may suggest that the network needs more information or features in the input to separate the images in the correct methods. Another

possibility is to use other classifiers as the support vector machine (SVM).

V. Conclusion

This paper presented a study of bottom-up features of human focus in natural images. 21 features with 9 methods of prediction were analyzed and compared to human notes to 1003 images of the MIT1003 database.

Experiments were conducted to study the Visual Attention Model under 4 main questions: (i) How the features present in the image influence the efficiency of the saliency maps (SM)? (ii) What are the behaviors of the features on the ocular focus regions and in the rest of the image? (iii) How the individual behavior of the prediction methods acts as the complexity of the scene raises? (iv) Is it possible to analyze an image and indicate the best method to predict its focal points?

For the efficiency measurement we proposed the F factor, which measures the capacity of the focal regions of an image to be predicted by automated methods.

After sorting the images decreasingly in relation to the F factor, it is noticed that the studied database has 3 groups of images. The first group has a high F factor; the second has an average F factor; and the third group has a low F factor. The main observation on this database is that the average F factor group is much bigger than the others.

This sorting technique was used to analyze several factors. It was noticed that features as entropy, intensity standard deviation and colors on the channels H and V have a strong relation with the F factor. A particular case is observed to the channel H of the HSV system. In this case, the smaller the dispersion on this channel is, the bigger the F factor is. Its average, on the other hand, is the feature that less relates to the F factor. This indicates that smaller diversification of colors helps the prediction.

Also, on the studied features, the dispersion measures tend to be more related to the F factor, while the centrality measures tend to have less relation to the F factor. This suggests that, the bigger the variety of information on the image, the more complex the predictions are to the automated methods.

Besides, the spatial features, as the number of disconnected regions, have a strong relation with the F factor. Assuming that the number of regions relate to the number of objects on the scene, having fewer objects on the scene seems to make the prediction easier.

This paper presented an evaluation of region features where the ocular focus happened and in the rest of the image separately.

With an independent analysis, for the MIT1003 database, the humans visualized the regions that had the brightest colors.

Furthermore, there are visible differences on the entropy value in foreground and background regions. The background entropy tends to be bigger than the focused region. This suggests that even in images that have a high volume of information, the humans can distinguish the background and focus on the regions of interest. Despite the entropy, the same patterns can be noticed on the standard deviation on all 3 channels of the HSV system.

However, as the difference between the deviations and the entropy grows, the less efficient the automated methods are. This suggests that other factors can influence the capacity to detect the attention points, like the spatial features, or scenes that have more than one region of interest. Because of this, it is possible to notice that as the number of focused images grows, the less efficient the prediction methods are.

This behavior suggests a relation between the size of the focused area and the number of regions, once the human eye passes through a bigger area while interpreting the scene.

Considering these behaviors, is possible that newest methods uses the region of interest to have better efficiency on the Saliency Map.

On the analysis of the saliency methods models, each image of the MIT1003 database was grouped according to the method that best created a saliency map to the respective image. The studied features were compared among the groups of each method.

On this comparison, it can be noticed that the centrality method reaches high efficiencies to images with higher entropy. On the other hand, it can be noticed that the methods without the centrality model lose efficiency on more complex images, they can even be worse than the centrality model.

All these factors suggest that, to images that are hard to predict the focus, the best strategy is to indicate the center of the image as the focus region. This fact can take advantage of the tendency of images being centralized on the region of interest.

With the intention to have an indication of which method is better to which image, on this paper a neural network was created, it has as input the features within the scene and as output the indication of one of the 9 methods used in this paper.

Evaluating the accuracy of the network, it was noticed that more than half of the indications do not select the method with the bigger efficiency. Besides, there is a dense concentration of indications to the method GBVS.

Therefore, this behavior points that was not possible to establish a separation hyperplane between the methods that would generate the best SM to a specific image. So, new features or classifiers can be considered to achieve an efficient classification, with higher accuracy.

This paper proposed a methodology to analyze which features influence the focus and the efficiency of the methods, and how they do it. As result of this analysis, behaviors that can be used on next papers were found. For example, on new prediction methods or a more efficient selection of method.

It is possible to explore new prediction models that use the patterns found on this paper. Thus, techniques that maximize the difference of entropy and intensity of the channel H and V on the saliency regions of the image can be developed. Besides, these techniques must grant bigger saturation level on the attention points and the minimum of disconnected regions.

In this paper we also noticed that, on an image with high predict capacity, the centrality model harms the saliency map. However, on more complex images the centrality model helps on the prediction. This way, new researches that propose an adaptive selection of weights on the model may have promising results.

New automated method selectors can be explored, in order to reach better performance. In order to achieve good results, there are 3 possible approaches: the first one, expanding the input features, using high level features also, as face detection; the second approach is using others classifiers, for instance the SVM (support vector machine); and the third is the possibility to study the patterns and develop a method selection through the features found on the saliency maps, and not only on the original images.

REFERENCES

[1] L. Elazary, L. Itti, Interesting objects are visually salient, Journal of Vision 8 (2008) 3,1–15.

[2] J. M. Henderson, J. R. Brockmole, M. S. Castelhano, M. Mack, Chapter 25 - visual saliency doesnot account for eye movements during visual search in real-world scenes, in: R. P. V. Gompel, M. H.Fischer, W. S. Murray, R. L. Hill (Eds.), Eye Movements, Elsevier, Oxford, 2007, pp. 537 – III.

[3] N. Ouerhani, Visual attention: from bio-inspired modeling to real-time implementation, Universit´e deNeuchˆatel, 2004.

[4] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention,Vision research 42 (2002) 107–123.

[5] A. Borji, L. Itti, State-of-the-art in visual attention modeling, Pattern Analysis and Machine Intelli-gence, IEEE Transactions on 35 (2013) 185–207.

[6] C. M. Privitera, L. W. Stark, Algorithms for defining visual regions-of-interest: Comparison with eyefixations, Pattern Analysis and Machine Intelligence, IEEE Transactions on 22 (2000) 970–982.

[7] A. Mishra, Y. Aloimonos, C. L. Fah, Active segmentation with fixation, in: Computer Vision, 2009IEEE 12th International Conference on, IEEE, pp. 468–475.

[8] D. DeCarlo, A. Santella, Stylization and abstraction of photographs, in: ACM Transactions on Graphics(TOG), volume 21, ACM, pp. 769–776.

[9] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applicationsin image and video compression, Image Processing, IEEE Transactions on 19 (2010) 185–198.

[10] R. Rosenholtz, A. Dorai, R. Freeman, Do predictions of visual perception aid design?, ACM Transac-tions on Applied Perception (TAP) 8 (2011) 12.

[11] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict humanfixations, in: MIT Technical Report.

[12] H.-C. Nothdurft, Salience from feature contrast: additivity across dimensions, Vision Research 40(2000) 1183 – 1201.

[13] L. Itti, C. Koch, Computational modelling of visual attention, Nature Reviews Neuroscience 2 (2001)194–203.

[14] G. A. W. Lopes, Reconhecimento de Objetos Utilizando Percep¸c˜ao Multissensorial Competitiva Baseadaem Redes Complexas, Thesis, Centro Universit´ario FEI, 2016.

[15] A. L. Yarbus, Eye Movements and Vision, Plenum. New York., 1967.

[16] M. K¨ummerer, L. Theis, M. Bethge, Deep Gaze I: Boosting Saliency Prediction with Feature MapsTrained on ImageNet, ArXiv e-prints (2014).

[17] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances In Neural InformationProcessing Systems 19, MIT Press, 2007, pp. 545–552.

[18] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans. Pattern Analysisand Machine Intelligence 34 (2012) 1915–1926.

[19] F. L´opez-Garc´ıa, R. Dosil, X. Pardo, X. Fdez-Vidal, Scene Recognition Through Visual Attentionand Image Features: A Comparison Between SIFT and SURF Approaches, INTECH Open AccessPublisher, 2011.

[20] A. Garcia-Diaz, V. Leborn, X. R. Fdez-Vidal, X. M. Pardo, On the relationship between opticalvariability, visual saliency, and eye fixations: A computational approach, Journal of Vision 12 (2012)17.

[21] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEETrans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259.

[22] G. Kootstra, N. Bergstr¨om, D. Kragic, Using symmetry to select fixation points for segmentation, in:ICPR, IEEE Computer Society, 2010, pp. 3894–3897.

[23] M. Cerf, J. Harel, W. Einhaeuser, C. Koch, Predicting human gaze using low-level saliency

combinedwith face detection, in: J. C. Platt, D. Koller, Y. Singer, S. T. Roweis (Eds.), Advances in NeuralInformation Processing Systems 20, Curran Associates, Inc., 2008, pp. 241–248.

[24] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vision 45 (2001) 83–105.

[25] J. Li, Y. Tian, T. Huang, W. Gao, Probabilistic multi-task learning for visual saliency estimation invideo, International Journal of Computer Vision 90 (2010) 150–165.

[26] S. J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Pearson Education, 2 edition,2003.

[27] M. de Brecht, J. Saiki, A neural network implementation of a saliency map model, Neural Networks19 (2006) 1467 – 1474.

[28] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention,Vision Research 40 (2000) 1489–1506.

[29] D. M. Green, J. A. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966.

[30] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Y. Weiss, P. B. Sch¨olkopf, J. C.Platt (Eds.), Advances in Neural Information Processing Systems 18, MIT Press, 2006, pp. 155–162.

[31] R. J. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images,Vision Research 45 (2005) 2397 – 2416.

[32] T. Judd, Understanding and predicting where people look in images, Ph.D. thesis, MassachusettsInstitute of Technology, Cambridge, MA, USA, 2011.

[33] M. Mancas, O. L. Meur, Memorability of natural scenes: The role of attention, in: IEEE InternationalConference on Image Processing, ICIP 2013, Melbourne, Australia, September 15-18, 2013, IEEE, 2013,pp. 196–200.

[34] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEEInternational Conference on Computer Vision (ICCV).36