THE APPLICATION OF HUMAN VISUAL ATTENTION IN MACHINE VISION

Mohammad A. N. Al-Azawi[1,2]

[1]Center for Computational Intellegence, De Montfort University, Leicester, UK

[2]Dept. of Computer Science, Oman College of Management and Technology, Barka, Oman

E-mail: mohd.alazawi@omancollege.edu.om

KEYWORDS

Human Attention, Computer Vision, Image, Neural Net.

ABSTRACT

Machine vision is still a challenging topic and attracts many researchers to research in this field. The main difference between machine vision and human vision is that, machine can see images as a set of pixels, while human can used cognitive capabilities in identifying the contents of the image. Attention is one of the important properties of Human Vision System, with which the human can focus only on part of the scene at a time; scenes with more abrupt features shall attract human attention more than other regions. In this paper, we shall simulate the human attention and discuss its application in machine vision and how it will improve the result of the retrieval process and image identification and understanding. Artificial intelligence shall be used in the proposed algorithm to cluster the salient points that were obtained either from eye trackers or from saliency extraction techniques.

INTRODUCTION

Machine vision is still one of the challenging topics that attracts the researchers to examine and develop a reliable Machine Vision Systems MVS. The main problem of MVS is that, machine can see the scenes as a set of pixels which cannot recognize or understand them. In MVS, the machine depends on some features (descriptors) to describe the contents of an image, usually these descriptors can be extracted from features such as colour, texture, and shape. In most image identification systems, the machine recognizes the image as a whole, which does not always give acceptable results. This is because different images may have similar descriptors such as having similar colour distribution (histogram). In order to overcome such kind of problems identification by parts can be used, in which the image is divided into connected regions (segments) and then these segments are identified. This process may improve the identification results but it is still suffering from identifying unimportant regions in the image. The interest in developing systems enspired by Human Vision Systems (HVS) has been increased recently, in which, the researchers are trying to give the machine some properties from HVS. Attention is one of the most important properties of HVS, which improves the identification process. In this paper, we shall introduce a new method that utilises the human attention principle to improve the abilities of MVS. Artificial intelligence has been used here to improve the speed and performance of the proposed algorithm. The paper includes a brief discussion of the human attention phases and function, and how to simulate them in MVS.

HUMAN ATTENTION AND IMAGE SALIENCY

This section presents a brief description of the main features, aspects, and phases of human attention, in addition, it discusses the definition of saliency in images and the main techniques that were proposed to extract the salient regions from images.

Human Attention

Human Attention is the process of selecting a subset of the available information upon which to focus. Mainly it has three aspects and goes through these aspects in sequence (Ward, 2008):

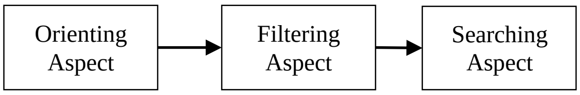

Figure 1 Attention Phases (Aspects)

In Orienting Aspect, when a human receives more than one stimulus, he orient his attention toward only one of them to focus on. Several types of orientation were proposed such as, orienting reflex, goal-driven, and stimulus-driven orienting. In the orienting reflex, pop-out actions can capture human attention, such as, a large object and with different colour in a smooth background. While in the goal-driven orientation attention is oriented to a location in space or to an object in a goal-driven manner, often based on a cue that tells where to look.

When attention is captured by some object, oriented to a specific location, and then moved to another location it will not return to the previous location directly or it is inhibited to orient or returning to the original location for a period of time; this is known as “Inhibition of Return” (Ivanoff & Klein, 2008) (Itti et al. 1998).

In filtering Aspect, which comes after orientation, the attention acts as a filter to remove information with less importance and focus on information with higher important. For instance, in an image with abrupt subject, like a bird in a sky, the human may see the bird while the sky will be blurred and not seen by him. From this, it is clear that the brain has filtered the image into two parts, one of them is with higher importance, which is the bird, and the second one is with less importance, which is, the sky in this example. The brain filtered out the sky and the bird will captured all the attention. Human has no ability to divide his attention among more than one stimulus, and the above example clarifies this idea, because the human cannot see the sky and the bird at the same time, we shall call this as singularity of attention (SOA).

The search phase comes as the last phase in which, when someone knows what he wants to find but does not know where to find it in the image, attention will be involved in this phase. Search can be classified as easy versus difficult search; also, it can be classified as automatic versus controlled search.

In easy search, (also known as pop-out, or parallel search), when an observer searches a well-recognized item in a set of items, it will be easy for him to find the required item regardless of the number of other items. The relation between the number of items and the time required to find the target will be a straight line with constant value as given in equation (1). In this function τ is the time required to find the target, n is the number of items in the image, and α is a constant number.

$$\tau = f(n) ; \tau = \alpha \qquad \textbf{(1)}$$

As shown in Figure 2, which shows a salient object in a set of similar items, regardless of the number of items, small number of items in (a) and large number of items in (b), the time required to find the red item is constant.

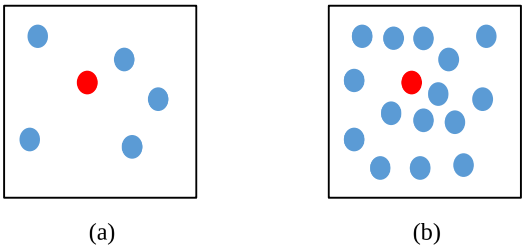

Figure 2: Easy Search (a) A Salient Item in Small Number of Items (b) A Salient Item in Large Number of Items

In difficult search, (which is also known as serial search); the searcher needs to find a particular item based on conjunction features. The search will be slow and the required time to find the target will increase linearly with the number of items in the image because one needs to focus on each item for a period of time. The relationship between the time required and the number of items is given in equation (2).

$$\tau = f(n) ; \tau = \alpha n \qquad \textbf{(2)}$$

An example about the search for conjunction features of an item is given in Figure 3, finding the red oval is very much easier in (a) than (b). Since there are more than one feature one needs to find (colour and shape in this case); one needs to look at each item and decide whether it is the target or not.

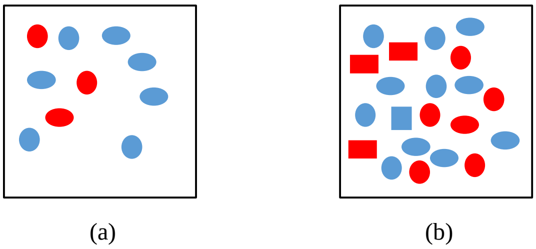

Figure 3 Difficult Search. (a) A Salient Item in Small Number of Items (b) A Salient Item in Large Number of Items

Search can be classified further into Controlled and Automatic; controlled search is usually used with difficult search. In order to achieve better search results, the same person may be requested to perform the search process. This process will train the searcher and hence, the search will become an automatic search. Automatic search is very much faster than controlled search, thus, training a searcher to perform some searching task will speed up and improve the results dramatically.

Pre-Attentive and Attentive Aspects

Pre-attentive processing is the unconscious accumulation of information from the environment. All information is considered as pre-attentive at the beginning then the brain process this information to filter it and give importance to each item based on the saliency of that item. Information with high saliency is more opt to be undergone under further attentive processing.

Some information proceeds from pre-attentive to attentive based on the saliency of the information and the goals and intentions of the observer. Many properties may give an item in an image the importance to pop-out from other items (destructors), such as, colour, shape, orientation, length, closure, size, curvature, density, contrast, hue, and others.

Pure vs. Cognitive Capture

The pure capture model uses the saliency of the stimulus and how it differs from the background. This model also known as bottom-up since it focuses on the visual saliency more than the conscious goal.

In contrast to pure capture, cognitive capture (also known as top-down selection) emphasize the idea that the searcher has an idea about the stimulus he searches for, which means that goals and intentions are presented in the search process. The brain pays attention to stimuli that have features to the target item.

Salient Regions Extraction

Salient regions are regions that attract the attention of a human, many reasons may attract human attention such as colour contrast, size, etc. Saliency extraction can be divided mainly into Bottom-Up Feature-Based Approaches (BUFB) and Top-Down Knowledge-Based Methods (TDKB) (Kapsalas et al. 2008). In BUFB, the salient regions are extracted based on the difference in features of the regions from other parts of the image. This difference can be measured based in some low level features that can be extracted from the image such as colour and texture. The second approach (TDKB), aims to construct the knowledge first, then search for interesting points according to the knowledge database. Another broad classification of the salient points extraction techniques was proposed by Toet (Toet, 2011), he classified the techniques as biologically-based, purely computational, or a combination of both.

Many publications have been published related to this issue; in this section we will highlight and study the significant publications and proposed methods.

Wavelet-Based Techniques

Wavelet is a multiresolution representation that expresses image variations at different scales, which gives information about the variations in the signal at different scales. It was used by several authors in which they have applied the principles of wavelet transform to extract the salient points (Loupias et al. 2000) (Tian et al. 2001) (Song et al. 2006) (Lin and Yang, 2007) (Arivazhagan and Shebiah, 2009). Loupias *et al.* has suggested the use of orthogonal Haar wavelet in extracting the salient points in an image (Loupias et al. 2000). They have considered high wavelet coefficient at a coarse resolution corresponds to a region with high global variations. In their method, they find the variation at lower level and track this variation to the original image size, this tracking is performed form one level to the upper one. Song et al. have proposed the use of wavelet to identify the salient points in colour image; they have converted the colour image into its three bands, Red, Green, and Blue. The wavelet is calculated for each band from which the salient points can be extracted (Song et al. 2006).

The use of wavelet may give good results in a non-homogeneous image, like the example that was used in (Loupias et al. 2000), which was the cameraman example, while the obtained results are not that good in the case of images with mixed texture.

Location-Based Saliency

Kim, Park, & Kim in 2003 have used the central position and colour contrast as the features that give importance to the object (Kim et al. 2003). They assumed that photographer always focuses to put the important object in the middle of the image.

Corner-Based Techniques

Geometric features such as corners were used to identify the points’ saliency. Comer detectors are constant with image scaling, shifting, and rotation. Corners were considered as a measure of saliency firstly by Schmid and Mohr in 1997 as part of their effort to identify interest points locally (Schmid & Mohr, 1997) (Kapsalas et al. 2008). Several corner detectors can be used such as Harris and the modified approaches such as Moravec and SUZAN. In their paper published in 2000 (Loupias et al. 2000), Loupias et al. have criticized the use of corner-based techniques and showed the limitations of using this method. The limitations they presented are; first, important visual features are not necessarily to be corners, and second, corners may gather in small regions, like in texture images.

Feature Maps-Based Saliency

Early work in this field was done by Koch and Ullman (Koch & Ullman, 1985) and Itti *et al* (Itti et al. 1998), feature maps which are the maps of some features that are extracted from the image can identify the saliency of a point or a region based on extracting some low level features, such as, colour, intensity, texture, orientation, etc.

Itti *et al.* have developed a model to extract the regions of attention; which is considered as one of the most popular models in this field. In their model, they used image hierarchy and they used intensity, colour, and orientation features to distinguish the salient regions. The features were calculated by a set of centre surround operations. 42 feature maps were generated from these features, the saliency map is extracted by combining these maps. The inhibition of return was used here in order to prohibit the algorithm from considering the same salient object more than once.

The main drawback of this technique that it uses so many feature maps (42), in additional to the use of image pyramid, these two drawbacks may affect the speed of the algorithm. The second important drawback is that it extracts the interest regions sequentially, which means that one region at each iteration using winners-takes-all paradigm, which increases the computation time.

Frequency Spectra-Based Saliency

Frequency domain was used to extract the saliency of the objects in many literature such as (Bruce et al. 2007), (Hou & Zhang, 2007), (Li et al. 2007), (Achanta et al. 2009), (Zhou et al. 2010), and (Fang, et al., 2012). The idea behind using the frequency domain is that, they considered that salient points are usually of high change in frequency domain both in magnitude and in orientation. Bruce et al. (Bruce, et al , 2007) have suggested the use of magnitude to extract the salient regions in an image. They have divided the image into sub-images and used Fourier Transform to convert the sub-image to the frequency domain. After which they have calculated the magnitude of the image from the real and imaginary parts of the spectral image and consider the regions with high frequency as a salient region. In ref (Hou & Zhang, 2007), the authors also have considered the magnitude of the spectral image to extract the saliency map. They have considered that the image contains two parts, prior knowledge and innovation; they have considered the prior knowledge as the redundant information in the image and the innovation as the salient information. The saliency is extracting by taking the Fourier transform for the image, extract the log spectrum, then subtract the redundant part, and finally convert it back to spatial domain to get the saliency map. The main limitation of such techniques is that, they are considering regions with high frequency as the salient regions. Region of high frequency are regions with high local changes such as edges, and edges not necessarily to be salient regions.

Fixations and Saccades

In human vision behaviour studies, these two properties are important because they can give good impression about the human interest, which in its turn, will give good impression about the importance of the regions in an image. Fixation can be defines as the point at which the human may gaze for a period of time, while saccade is the fast movement of the eyes from one point to another in the image. Human may fixate his eyes on attracting regions, the most important points may be found at that location, while moving his eyes very fast from one point to another will generate the saccade regions. These regions maybe considered as unimportant regions, or, regions with less attention capturing.

Humans have a collection of passive mechanisms to reduce the amount of incoming visual information, i.e. the signal stemming from the photoreceptors is compressed by a factor of about 130:1, before it is transmitted to the visual cortex (Le Meur et al. 2006). The HVS has an active selection, involving eye movement. A saccade is a rapid eye movement allowing jump from one location to another. The purpose of this movement, which occurs up to three times per second, is to direct a small part of HVS field into the fovea to achieve a closer inspection. The last step corresponds to a fixation.

NEURAL NETWORK BASED GAZE POINTS CLUSTERING

Neural Nets can be used for identifying the regions of interest and the unimportant regions using the data obtained either from the eye trackers or from the salient points extraction algorithms. The neural networks need to be trained first using a set of images, and then they can be used for identifications. The image is divided into sub-image; every sub-image should be identified using the trained neural network.

Back propagation neural network as a classifier

The back-propagation training requires a neural net of feed-forward. Since it is a supervised training algorithm, both the input and the desired output patterns are required. For a given input pattern, the output vector is estimated by forward pass then an error value is calculated and propagated back through the network to update the weights of the network.

The neural net given in Figure 4 shall be used to perform the classification process, the network represent a multilayer perceptron, which is a feed forward network. The network consists of more than two layers i.e. it consists of one hidden layers between the input and output layers (Lendaris & Mathia, 1996).

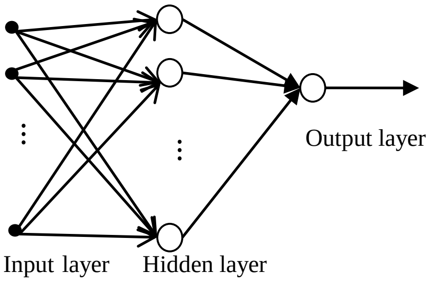

Figure 4 Multilayer Perceptron

As shown in Figure 4 the input layer is not connected directly to the output layer but it is connected through the hidden layer, the output vector of the hidden layer is used as input to the output layer. With multilayer perceptron more complex problems can be solved which are not possible to be solved with single layer. Multilayer can separate the space to more complex decision regions than single layer; every layer is trained in the same training algorithm of single layer perceptron. The number of nodes in each layer is different from layer to another.

The input to the neural network will be the set of descriptions of the sub-images surrounding the gaze and saccade points. Let us assume that the set of gaze points or the points of interest is $\boldsymbol{S}$ and the set of unimportant points or the saccades is $\boldsymbol{U}$, then the set of all points that are used to train the NN will be the union of these two sets to form the set $\boldsymbol{P}$. i.e.

$$\begin{aligned} &\boldsymbol{P} = \boldsymbol{S} \cup \boldsymbol{U} \\ &\boldsymbol{P} \subset \boldsymbol{I}, \text{ I is the set of all points in an Image} \\ &n_p = |\boldsymbol{P}|\,, n_s = |\boldsymbol{S}|, n_u = |\boldsymbol{U}| \\ &n_p = n_s + n_u \end{aligned} \tag{3}$$

Where n_p, n_s and n_u are the cardinalities of the sets $\boldsymbol{P}, \boldsymbol{S},$ and $\boldsymbol{U}$ respectively. We shall define a mapping from the set $\boldsymbol{P}$ to the set of real number that extracts the measures that are used in identifying the sub-images, i.e.

$$\begin{aligned} &\rho: \boldsymbol{P} \to \boldsymbol{R}^n \\ &\hat{f} = \rho(p) \; \forall \, p \in \boldsymbol{P} \end{aligned} \tag{4}$$

Where $\hat{f}$ is a vector of features that are extracted from the sub-image surrounding the points p. This vector of features will be used in training the neural network to identify the rest of regions.

The input to the neural network shall contain two different types of data;

1. The vectors of features that are extracted from the region surrounding the interest points $\hat{f}^s$.
2. The vectors of features that are extracted from the region surrounding the unimportant points $\hat{f}^u$.

Since the training of such kinds of NNs is supervised learning, which means that we need to provide the NN with both input and desired output vectors then the NN will have only one output to specify the saliency level of the rest of the sub-images. The output vector $\hat{y}$ will have real value between 0 and 1. During the training phase, it will be given 1 for the important regions and 0 for the unimportant regions.

For the purpose described above, we shall use the back propagation training technique with number of input nodes equals to the number of features that are used in describing or identifying the sub-image. The number of nodes in the hidden layer is equal to one half of the number of nodes in the input layer, this number is reasonable to produce the required nonlinearity of the network, finally, the network will have one output to identify the saliency level. The bipolar sigmoid function $g(x) = \frac{1-e^{-\alpha x}}{1+ e^{-\alpha x}}$ shall be used as an activation function, with α is a tuning factor that may affect the slop and the convergence of the function. Bipolar sigmoid was selected instead of the sigmoid function to increase the range of the output values between -1 and 1, while it is between 0 and 1 in the regular sigmoid function.

The features that have been used in this test are the histogram features of the three color bands, other features can be used as well.

Figure 5 shows the result of applying the neural network in identifying the important regions from important points. As shown in the figure, the red circles represent the regions surrounding the important points and the blue circles represent the regions surrounding the extracted saccade points. The sub-images are of size of 7 × 7. The neural network was trained using the set of regions given in Figure 5 (b) and then the trained neural network was used to identify all other sub-images with overlapping of six pixels between them, as shown in (c) which shows the map of the neural network output, and (d) which shows the identified important regions.

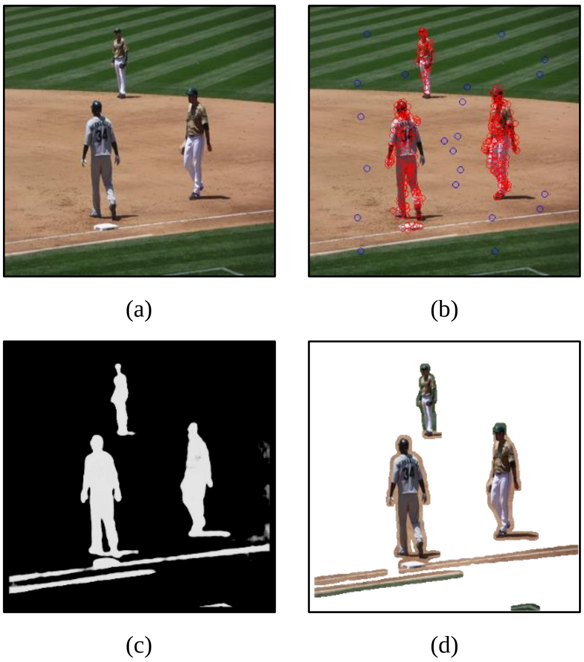

Figure 5 The application of neural network on salient regions extraction, (a) original image, (b) important points (red) and saccade points (blue), (c) saliency map, (d) the extracted objects

The same trained neural network can be used to extract the salient region in similar images as shown in Figure 6.

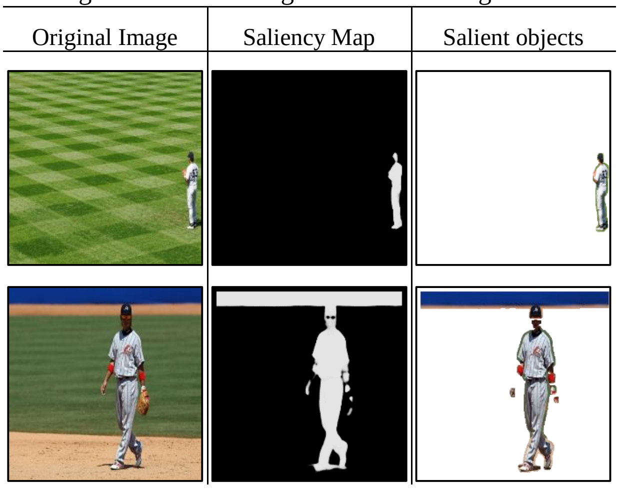

Figure 6 The application of the trained neural network on salient regions extraction on images similar to the image in Figure 5

CONCLUSION

In this paper, we have utilized the principle of human attention in HVS to introduce a new important regions identification algorithm that specifies the important regions in an image. As it was discussed, large amount of data can be reduced and ignored such as the background. The proposed technique gave excellent results and it worked in a very good way when applied on the saliency datasets. The method can be used in different applications such as image retrieval systems, computer vision, image and video compression, and others. The use of NN has improved the performance of the proposed algorithm since with conventional correlation techniques, the similarity among the sub-images should be calculated for each sub-image with all other sub-images, which is a computationally inefficient method. With NN, the comparison was very much easier parallel, which reduced the amount of calculations required.

REFFERENCES

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned Salient Region Detection. *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009.*

Arivazhagan, S., & Shebiah, R. (2009). Object Recognition Using Wavelet Based Salient Points. *The Open Signal Processing Journal, 2*, 14-20.

Bruce, N., Loach, D., & Tsotsos, J. (2007). Visual Correlates of Fixation Selection: A Look at the Spatial Frequency Domain. *IEEE International Conference on Image Processing. ICIP 2007 (Vol 3).* San Antonio, TX.

Fang, Y., Lin, W., Lee, B.-S., Lau, C.-T., Chen, Z., & Lin, C.-W. (2012). Bottom-Up Saliency Detection Model Based on Human Visual Sensitivity and Amplitude Spectrum. *IEEE TRANSACTIONS ON MULTIMEDIA, 14*(1), 187-198.

Hou, X., & Zhang, L. (2007). Saliency Detection: A Spectral Residual Approach. *IEEE Conference on Computer Vision and Pattern Recognition. CVPR '07.* Minneapolis, MN.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254-1259.

Ivanoff, R. M., & Klein, J. (2008). Inhibition of return. *Scholarpedia, 3*(10), 3650. Retrieved from http://www.scholarpedia.org/article/Inhibition_of_return

Kapsalas, P., Rapantzikos, K., Sofou, A., & Avrithis, Y. (2008). Regions of interest for accurate object detection. *International Workshop on Content-Based Multimedia Indexing, CBMI 2008.*, (pp. 147-154). London.

Kim, S., Park, S., & Kim, M. (2003). Central object extraction for object-based image retrieval, CIVR'03. *Proceedings of the 2nd international conference on Image and video 03*, (pp. 39-49). Heidelberg .

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

Le Meur, O., Callet, P. L., Barba, D., & Thoreau, D. (May 2006). A Coherent Computational Approach to Model Bottom-Up Visual Attention. *IEEE Trans. Pattern Analysis and Machine Intelligence, 28*(5), 802-817.

Lendaris, G., & Mathia, K. (1996). Efficient Numerical Inversion Using Multilayer Feedforward Neural Networks. *World Congress on Neural Networks (WCNN'96).* San Diego, California.

Li, J., Levine, M., An, X., Xu, X., & He, H. (2007). Visual Saliency Based on Scale-Space Analysis in the Frequency Domain. *JOURNAL OF LATEX CLASS FILES, 6*(1), 1-16.

Lin, D.-W., & Yang, S.-H. (2007). Wavelet-Based Salient Region Extraction. In *Advances in Multimedia Information Processing – PCM 2007. Vol. 4810* (pp. 389-392). Honk Kong: Springer.

Loupias, E., Sebe, N., Bres, S., & Jolion, J.-M. (2000). Wavelet-based salient points for image retrieval. *Proceedings of International Conference on Image Processing*, (pp. 518-521). Vancouver, BC, Canada.

Schmid , C., & Mohr, R. (1997). Local greyvalue invariants for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence, 19*(5), 530-535.

Song, H., Li, B., & Zhang, L. (2006). Color Salient Points Detection Using Wavelet. *Proceedings of the 6th World Congress on Intelligent Control and Automation.* Dalian, China.

Tian, Q., Sebe, N., Lew, M., Loupias, E., & Huang, T. (Oct 01, 2001). Image retrieval using wavelet-based salient points. *J. Electron. Imaging, 10*(4), 835-849.

Toet, A. (2011). Computational versus Psychophysical Bottom-Up Image Saliency: A Comparative Evaluation Study. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 33*(11), 2131-2146.

Ward, L. M. (2008). Attention. *Scholarpedia, 3*(10), 1538. Retrieved 9 5, 2012, from Scholarpedia: http://www.scholarpedia.org/article/Attention

Zhou, B., Hou, X., & Zhang, L. (2010). A phase discrepancy analysis of object motion. *Proceedings of the 10th Asian conference on Computer vision ACCV'10,* (pp. 225-238).