# Biologically-Inspired Face Detection: Non-Brute-Force-Search Approach

Christian Siagian

Department of Computer Science
University of Southern California
Los Angeles, CA 90089

Laurent Itti

Department of Computer Science
University of Southern California
Los Angeles, CA 90089

## Abstract

*We present a biologically-inspired face detection system. The system applies notions such as saliency, gist, and gaze to localize a face without performing blind spatial search. The saliency model consists of highly parallel low-level computations that operate in domains such as intensity, orientation, and color. It is used to direct attention to a set of conspicuous locations in an image as starting points. The gist model, computed in parallel with the saliency model, estimates holistic image characteristics such as dominant contours and magnitude in high and low spatial frequency bands. We are limiting its use to predicting the likely head size based on the entire scene. Also, instead of identifying face as a single entity, this system performs detection by parts and uses spatial configuration constraints to be robust against occlusion and perspective.*

## 1. Introduction

Face detection and recognition are two interrelated tasks that are useful in many applications such as Human-Robot interactions [1]. Although it can be argued which of the two are more difficult, if a face detection module could supply the necessary parameters such as location, size and pose, a recognition process could be greatly simplified [23]. The bulk of current research [2] assumes that these features can only be found by using blind search. That is, sequentially shifting, rotating, and resizing a search window over the input image and testing whether the window at each instance contains a face. Here, we use saliency (a measure of conspicuity, or visual attractiveness, of any location in the scene [4]) and gist (a rapid but rough analysis of the entire scene yielding a small number of scene descriptors [20]) together to localize a face in a manner more in agreement with human behavior.

Saliency [6] is a term for a degree of conspicuousness that a feature exhibits in the presence of its neighbors. In this context we use the assumption that faces are quite distinct from their respective backgrounds, one that holds for many images that we have tested. Because saliency only utilizes simple cues in the visual cortex such as intensity, orientation, and color, it does not have to reason at the object recognition level. Thus it can be implemented in a highly parallel and computationally inexpensive fashion. A system by Itti [3] has been shown to reliably predict which regions in an image attract visual attention of humans and other primates, as demonstrated by a high correlation between model saliency and actual human eye movements in [5]. It employs image pyramids for each cue (or channel) and performs a center-surround operation between several pairs of scales to detect the degree of conspicuousness throughout the image [3]. The result from each channel is then weighted and linearly combined to create a saliency map, with the location that fires the strongest becomes the next attended location. The saliency model then inhibits that point so that it can shift its attention and attend to the next most salient point. It should be noted that we can tune the weight of the channels to look for a specific object.

Because salient features, by definition, are isolated occurrences (mostly as a part of the foreground), it discards most of the information that pertains to the image as a whole (the background for example). Gist, on the other hand, looks at an image in its entirety and deduces characteristics that cannot be produced by a single location. Holistic features, such as scale of objects in general, as well as the structure of object placement in an image can be quite useful if it can be obtained in an efficient manner; that is, without identifying objects in the image. Torralba and Oliva [11] use Discrete Fourier Transform (its amplitude spectrum) to capture the degrees of high versus low spatial frequency in an image. This attribute can be used to approximate size of a head as large when the majority of the Fourier amplitude spectrum is concentrated around low spatial frequencies (measured in 1/pixel), or small when high frequencies predominate. However, it is not an obvious bet that gist can be a reasonable predictor of head size. Indeed, gist characterizes the whole scene, not the head. Part of the present study is to test this hypothesis. It is also important to note that the computational complexity of both saliency and gist is much lower than that of detecting faces. In addition,

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

saliency and gist alone are not specific to faces.

Thus given some idea of the approximate size of a face and its whereabout, we can concentrate on a specific and limited-sized region. In the system presented here, the detection process is accomplished by identifying face parts and the configuration between them. Detection by parts has a built-in advantage of being robust against occlusion. Using the most salient location as a starting point we can move about its neighborhood, testing whether we have found a part and shift focus accordingly. Because we are limiting ourselves to finding faces, each part can help direct us to point toward the most promising direction so as to accumulate more information. This system uses several three-layer neural networks with backpropagation, each outputs a probability value of the likelihood that its respective part is detected.

In the event that a salient location is not a face, the subsequent recognition by parts will almost assuredly wander aimlessly about the region before stopping because of lack of progress. The system then proceeds to inhibit the region just searched before asking the saliency module to supply the next most salient point. A problem occurs when that salient point is within a close proximity to the face (reasonable enough to expect the system to successfully localize the face) although not on the face. Because of a lack of initial cues when attending off face locations, the gaze module can veer away from the face. This can become a bigger issue when a portion of the face is included in the inhibited region. In general, this process can be difficult because here we are trying to move about a neighborhood efficiently to identify the region without having the slightest idea what the region context is. The system may require an extended knowledge of where the face is with respect to body parts such as neck and torso.

### 1.1. Previous Work

There is an abundance of face detection approaches in the computer vision literature. Most of them assume that in order to find faces the process has to be done exhaustively. Approaches from simple template matching to neural networks have been quite successful in limited domains. Rowley, et al. [15] have implemented a planar rotation invariant neural network based face detection which uses a series of networks to first rotate (potential) faces to their upright position before hypothesizing the selected region. Later on, Schneiderman and Kanade[17] expand the domain to 3D (in-space rotation) using statistical methods in the form of a series of histograms. An approach by Turk and Pentland [21] utilizes Principal Component Analysis for feature extraction and come up with what they called eigen-faces to classify faces as well as identifying individuals. In addition work by Osuna, Freund, and Girosi [13] built a face finding system using a support vector machine.

## 2. Implementation

The following figure 1 illustrates the schematic of the system and where the different modules discussed above fits in together.

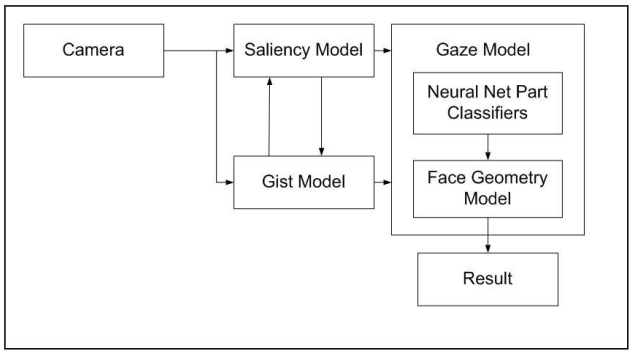


Figure 1: The architecture of the face detection system

### 2.1. Gist Model

For gist, we implement a sub-system adapted from recent work by Torralba and Oliva [11] by applying a Discrete Fourier Transform to the whole image. Discrete Fourier Transform can provide information such as dominant contour, orientation, and frequency. Given that the information is holistic, the dimensions of the image can be decimated to reduce amount of data to be computed. Just by comparing the Fourier transform of two images we can predict a difference in high-level characteristics. Figure 2 is an example of the Fourier transform comparison of two images where the first one is contains predominantly higher frequency signal while the second one contains more of the lower frequency signal. For illustration purposes, the Fourier amplitude spectra represented below have been gamma-corrected to emphasize low-amplitude components.

To define a set of features describing the resulting Fourier Transforms, we apply a bank of log-Gabor filters to localize the activities of different ranges of frequency at various orientations. By performing a dot product between a single filter and the Fourier Transformed image we produce one feature. By the convolution theorem, convolution by a log-Gabor filter in the intensity domain is reduced to a point wise product of the Fourier-transformed image and the Fourier-transformed filter in the Fourier domain. Using five different frequencies at eight different orientations, forty features altogether are now available for classification. The following figure 3 shows how the filters look like in the Fourier domain.

The classification process is implemented using a series of three-layer neural networks with back propagation. The system keeps a pyramid (decimated by a power of two) of

Figure 2: images with their corresponding Fourier Transform

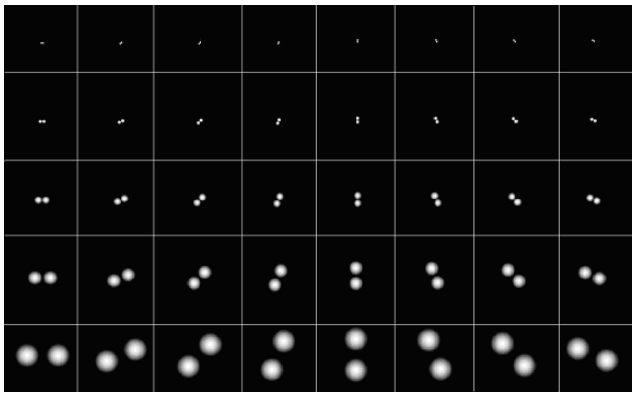

Figure 3: Fourier amplitude spectra of the log-Gabor Filters used to extract features from the Fourier Transform

the input image. The network considers the first five level and outputs of the likelihood a face is detected at each level. Figure 2 shows two different scale levels with the figure above at a lower level than the figure below because objects (including the faces) appear at higher resolution.

## 2.2. Saliency Model

As stated in the introduction, salient locations are identified by applying center-surround operation on various low-level cues in the visual cortex, weighted and summed together. Further details for the saliency model can be found in [3]. Since the input is limited to black and white images, the weights of the color and motion channel is set to zero. We do not tune the weight of the channels for faces for it does not seem to improve the results because face has quite a contrasting appearance from different point of views. The following images (Figure 4) show that the faces in the image are captured quite well using the saliency model without priming.

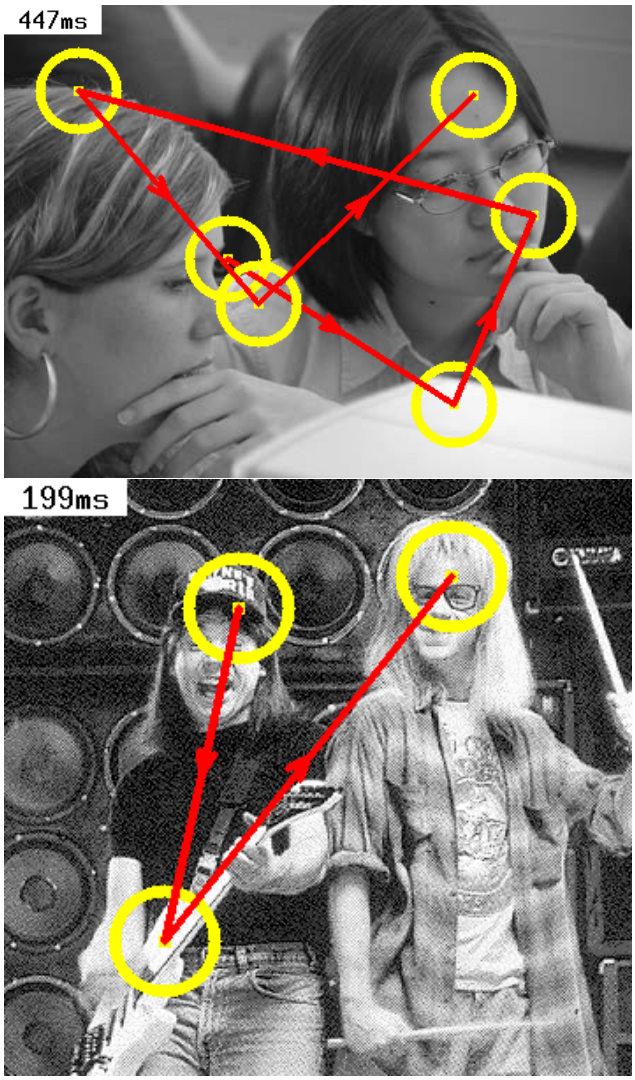


Figure 4: Trajectory of salient points of a few images

## 2.3. Part Detection

We use parts such as eyes, nose and mouth because, aside from being distinctive and stable, their identification only requires an edge cue. This is because when these parts are convolved they have the appearance of an edge. Studies from developmental vision show that infants demonstrate an innate predisposition towards face-like configurations (not necessarily a face) involving shapes resembling eyes, nose and mouth without the use of texture or color even within days of when they were born [18]. This suggests that those three parts may hold a special key to face detection. In addition, these face parts can point us to a specific direction to

find other parts as opposed to regions such as forehead or cheek, in which the next part can be attached in any direction in the case when the face orientation is unknown.

In practice, the part detectors are the one aspect that directly affects the performance of a face detection system. It is imperative that these modules give a reasonable estimate (probability) of whether a part has been found in a certain location. In our system, each part is detected using a three-layer neural networks with back propagation. Pattern matching using neural networks is easy to implement because all the designer needs to do is to find the target and distraction patterns. Some care is required to make sure all the border cases are included and overall coverage of the domain is accomplished. One difficulty with using neural networks is that whenever the network gives an incorrect classification, there is no way of knowing which way it is biased towards; the probability landscape is unexplainable. However, false positives are not so much a problem as false negatives because at each successive step the system will reevaluate the correctness of the previous estimate. Pattern such as faces [15], letters of the alphabet [7] and roads [14] have all been successfully detected by neural networks.

Illumination variation is a persistent problem in using neural networks classification, especially when using straight intensity information. Some preprocessing techniques such as histogram equalization [15] is usually applied to raw values. In our system, we subtract the image with its mean intensity and divide the result by its standard deviation. In addition, instead of using the intensity values, we use its Gabor filter outputs to be in agreement with the Visual Cortex. The orientation channel in the Saliency model, which utilizes Gabor filters to localize salient edges, provides these information. A stipulation about detection by parts: the face has to be big enough so that most parts are distinguishable. In addition, instead of inputting a whole image window, we select certain points within the window to reduce the input size to the network. This is similar to the set of landmark points within a deformable graph by von der Malsburg [25], although his work also involved accurately localizing individual landmarks using dynamic link matching. We experience that our results do not suffer by approximating the landmark point locations for it is able detect a part even when the window is not lined up correctly. Figures 5 shows a result when inputting each window in an image to each network. We use approximately a total of 200 negative and positive samples for each part. Note that in actual operation we only run the detectors on points suggested by the gaze module, about twenty points per attention shift.

## 2.4. Overall system

Putting it all together, the system starts by using the saliency module to direct itself to the most salient point of the image at the scale (in the image pyramid) estimated by the gist module. It then searches for face parts locally within a radius of the size of a face model. The system has a model that specifies the size of each part as well as their relationship. The size of the face model with the parts at the right configuration is 40 by 50 pixels. The parts itself has varying size from the smallest (eye) being 12 by 6 pixel to the largest (nose) being 14 by 20 pixels.

Figure 5: The resulting neural network based eye, nose and mouth detector

Once arriving at the most salient location, we select a few local salient points. Subsequently, when some parts are detected, additional context points are also added to the list. Context points are likely locations of parts given that a part is already confidently detected. Using a probabilistic grading, the next local attended point is the one that shows the most promise. By promise, the one that has the highest part probability multiplied by the probability of the connected part (found previously) and their geometrical relationship probability. Currently, the system only considers a binary relationship of two parts. This relationship can be expanded to ternary or even more complex grouping. The system will

Figure 6: Step by step detection of a face.

continue to shift focus until we do not make progress or a face is found in which case we can inhibit the region and find another face. Figure 6 illustrates the steps the system performed to discover a face. Note that this is the third most salient point (after the upper forehead and the shirt). One thing to notice is that the candidate points at each step are selected only with respect to the last attended point to minimize duplicate hypothesis. This aspect can also be improved by considering all the previous attention points as well as using a heuristic to improve local coverage.

# 3. Results

The straight-forward way to test the overall system is by using a series of unconstrained images of people and keeps track of the mistakes, false positive and negative. Currently, however, we have only tested the overall system with a series of isolated face images. Although the system remains true to the main promise that it is not doing a blind search. First off, we will test the system in details by evaluating the performances of each module. Afterwards we will assess the overall system.

## 3.1. Saliency Model

Table 1 displays the success percentage of the saliency model to locate a face in fifty randomly chosen images. The first three rows (total included) are a comparison between images of a lot (more than five) versus a few (less than five) people. The last three rows (total included) compare a set of images of big (with respect to the image size) versus small faces. These two tests uses the same set of images. Among the fifty images, there are 12 images with five or more people and 38 images of less than five people. Among the same set of images, there are 32 images with large faces and 18 images with small faces. After one (noted by a1 in Table 1) and five (noted by a5 in Table 1) means success in detecting a face after one and five attention shifts, respectively.

Table 1 shows that the saliency model is able to direct itself to a point near a human face (in less than five tries) 45 out of 50 times, while locating a spot on top of the face 39 out of 50 times. What is not shown in the table is that the saliency model is not an exhaustive locator. The more faces in the image the more it misses. This is because the more faces in the image the less salient individual faces become. Moreover the smaller the ratio of the head compared to the

| | On Face | | | Near Face | | |
|---|---|---|---|---|---|---|
| | a1 | a5 | Total | a1 | a5 | Total |
| $\geq 5$ people | 1 | 9 | 12 | 2 | 8 | 12 |
| $< 5$ people | 18 | 30 | 38 | 23 | 12 | 38 |
| Total | 19 | 39 | 50 | 25 | 45 | 50 |
| Large face | 16 | 27 | 32 | 19 | 31 | 32 |
| Small face | 3 | 12 | 18 | 4 | 14 | 18 |
| Total | 19 | 39 | 50 | 23 | 45 | 50 |

Table 1: Statistics of Experiments on Saliency Model

| Num Img | Misclass. | % | 2nd Misclass. | % |
|---|---|---|---|---|
| 120 | 43 | .37 | 26 | .22 |

Table 2: Testing Phase of Gist Module Neural Network

image size, the less likely it will be located.

### 3.2. Gist Model

Table 2 displays the success percentage of the gist model in predicting the head size given a vector of features from the images Fourier Transform. We are using 120 random images. The table shows two results, the number of misclassification when using only the prime scale, and when using the top two scales. We can see that the top two gives us a relatively good result (78 percent). There are a few sources of error. The first one is training sample not always perfectly labeled when the actual size of the head is in between two scale levels. The questioned head size is probably closer to one level but labeled as the other. The second source of error occurs mostly in larger head images. What happened is that the details of the face starts to show in which case the higher frequencies become more active and thus the networks misclassify it as a lower sized head.

### 3.3. Part Detectors

Table 3 shows the part detector testing session. It displays the success percentage of each part detector in detecting its respective face part. Since each neural network outputs a likelihood value, we cut off the classification at .5; anything above is a positive detection. It shows that the part detectors perform very well on the testing sample (less than five percent error). One stipulation is that the testing set contains the same set of people as in the training sample except they

| Part | Num | Pos. | Neg. | Err.% | Err. | FP | FN |
|---|---|---|---|---|---|---|---|
| Eye | 377 | 200 | 177 | .04 | 5 | 2 | 3 |
| Nose | 363 | 200 | 163 | .02 | 3 | 1 | 2 |
| Mouth | 378 | 200 | 178 | .02 | 2 | 1 | 1 |

Table 3: Testing Results of Part Detectors

| Num Img | Class. | FP | FN | Avg. Sal. | Avg gist |
|---|---|---|---|---|---|
| 100 | 65 | 7 | 28 | 4.2 | 1.7 |

Table 4: Face Detection Results

are taken at a different pose. The part detectors will work well in general because the inputs are Gabor outputs of images that are already normalized, and thus a particular shape (as long as it is reasonably sharp against a background) in any lighting will look approximately the same. In any case, the system is designed so that the part detector does not have to perform as well as shown above. Thus we believe that this will not be a problem for general face detection task.

### 3.4. Overall System

With various information avaliable (location and likelihood of parts, likelihood of face size), the culminating step would be putting it together and detect faces. At this point we have tested the system on 100 images of one person from various poses (from frontal to extemely rotated). Table 4 reports the details of the experiments. As we can see, the success rate is 65% with the average of 4.2 saliency points and 1.7 gist prediction needed to localize the faces. Observe figure 7 for some example of the detected faces.

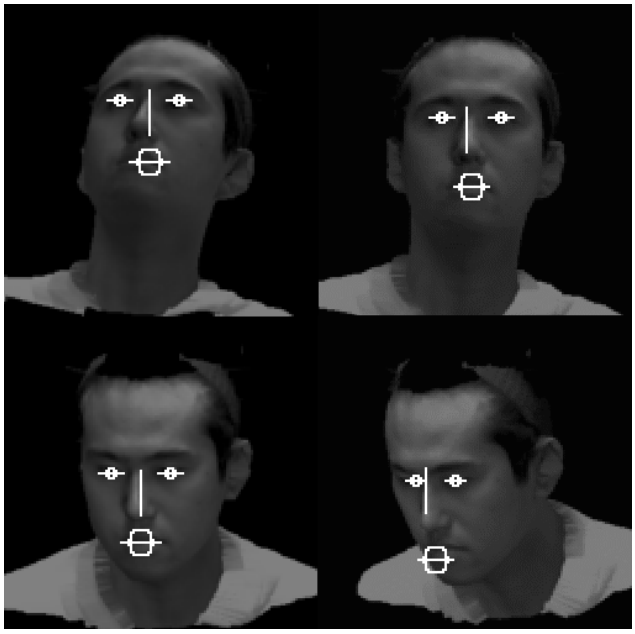

Figure 7: Examples of some of the detected faces

## 4. Discussion

The main trade-off in assessing the performance of the system is between time complexity versus accuracy and robust-

ness. In terms of time complexity, we put forth an argument that because the system do not reason in the object recognition level, all the computations are simple and highly operations (both in saliency and gist) and non-exhaustive (as in testing every possible location and scale in the image), it runs much faster than the regular brute-force-search approach.

In terms of accuracy and robustness, we observe that some of the reasons whya face cannot be localized. One is that the face itself is not salient or there are objects that more salient in the image. From the results, we observe that this only happens in less than 10 percent of the time. Also, if the faces are smaller than the model used (40 by 50 pixels), in which case the face parts are not recognizable in isolation, the system also fails. In its defense, the system is designed to find faces and not people. For a smaller sized person, the face is assumed because the head is attached to a human body. In this case we need to extend the system by adding a template of the whole face and a body. Given that the system already covers the remaining larger scale, the even smaller scale of these unclassified faces does not vary as much.

In addition there are more detailed aspects that can be improved. The system is adequate in detecting frontal-view faces, keeping in mind that the parts used are all from the same perspective. When we try to run the system on profile view faces, it does not work as well. This is because the parts do not reliably gives a good probability. It was our hope originally that the selected parts (especially the eye and mouth as they appear to be flat) can be identified from all angles in space using the trained network. A common solution is to create several classifiers from each view, spaced appropriately. The added problem would be that interpolation between views needs to be handled with care.

As for the gist model, we did not come in with a high expectation that it could perform well as a predictor of head size because gist itself characterizes the whole image, not the head. However, it performed above expectation. At this point the gist module takes in the whole image as one with the thinking that it is looking to extract holistic characteristics. We can, conceivably, divide the images into a grid or express the image with more features. For example, a better performance may be obtained by dividing the image into a four-by-four sub-region. Another problem faced by the gist module is the case where the image contains multiple persons in very contrasting depth. The current solution is to set the ground truth of the training set to classify the image as an image with multiple resolutions.

In the end we look back at the events when the salient points do not land directly on a face but nearby. This fact often causes the system to perform a random search with no sense of direction once it directs itself away from the nearby faces. There may be some learning that could be done or heuristics to implement so the local search has more purpose to speed up the exploration process. This would be among the first improvements on the current system.

# 5. Summary and Conclusions

In this paper we have successfully demonstrated that face detection does not have to be done in a brute-force search way. With the combination of isolated salient cues and holistic information, we can quickly focus on smaller region inside any images. As a result the system has a much faster running time.

It is also shown that the success of a system still comes down to the nitty-gritty templates (or neural network in this case). In order to detect a face from all perspective, the system has to be able to robustly detect eyes and other parts from different views. In addition, difficulties also comes in from the deforming configuration of the face parts in the three-dimension space.

# Acknowledgments

I would like to thank professor Christoph von der Malsburg and the Laboratory for Computational and Biological Vision for providing the series of face images used throughout the research.

# References

[1] Fong, T., I. Nourbakhsh, and K. Dautenhahn, "A Survey of Socially Interactive Robots," *Robotics and Autonomous Systems,* 42(3-4), pages 143-166, 2003.

[2] Forsyth, D., and J. Ponce, *Computer Vision A Modern Approach,* Prentice Hall, 2002.

[3] Itti, L., "Models of Bottom-Up and Top-Down Visual Attention," *California Institute of Technology. Ph.D. thesis,* Jan 2000.

[4] Itti, L., and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience,* Vol. 2, No. 3, pp. 194-203, Mar 2001.

[5] Itti, L., "Automatic Foveation for Video Compression Using a Neurobiological Model of Visual Attention," *IEEE Transactions on Image Processing,* in press.

[6] Koch, C., and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitery," *Hum Neuro-biol,* 4, pp. 219-27, 1985.

[7] Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner, "Gradient Based Learning Applied to Document Recognition," *Proceedings to IEEE 86(11),* 2278 - 2324. 1998.

[8] Matsugu M., K. Mori, Y. Mitari, and Y. Kaneda, "Subject Independent Facial Expression Recognition with Robust Face Detection Using a Convolutional Neural Network," *Neural Networks,* v.16 n.5-6, p.555-559, June 2003.

[9] Miau, F., C. Papageorgiou, and L. Itti, "Neuromorphic Algorithms for Computer Vision and Attention." *Proc. SPIE 46 Annual International Symposium on Optical Science and Technology,* Vol. 4479, pp. 12-23, Nov 2001.

[10] Navalpakkam, V., and L. Itti, "Modeling the Influence of Task on Attention," 2003.

[11] Oliva, A., and A. Torralba, "Modelling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int'l J. Computer Vision,* vol. 42, no. 3, pp. 145-175, 2001.

[12] Oliva, A., A. Torralba, M. Castelhano, and J. Henderson, "Top Down Control of Visual Attention in Object Detection," *Journal of Vision,* 3(9), 3a. 2003.

[13] Osuna, E., R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," *IEEE Conference on Computer Vision and Pattern Recognition,* 1997 pp. 130 - 136.

[14] Pomerleau, D., "Neural Network Perception for Mobile Robot Guidance," *Ph.D. thesis, School of Computer Science, Carnegie Mellon University,* 1992

[15] Rowley, H.A., S. Baluja, and T. Kanade., "Rotation Invariant Neural Network-Based Face Detection," *In Proc. of Computer Vision and Pattern Recognition,* pages 38–44, 1998.

[16] Rybak, I.A., V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, and N.A. Shevtsova, "A Model of Attention-guided Visual Perception and Attention," *Vision Research,* 38(2):387–400, 1998.

[17] Schneiderman H., and T. Kanade. "A Statistical Method for 3D Object Detection Applies to Faces and Cars," *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* pages 746-751, 2000.

[18] Slater, A., "The competent infant: Innate organization and early learning in infant visual perception," *Perceptual development: visual, auditory and speech perception in infancy,* Hove: Psychology Press 1998, pp 105-128.

[19] Torralba, A., and P. Sinha, "Statistical Context Priming for Object Detection," *IEEE Proc. Of Int. Conf in Comp. Vision,* 1:763-770, 2001.

[20] Torralba, A., "Modeling Global Scene Factors in Attention," *JOSA - A,* vol. 20, 7, 2003.

[21] Turk, M., and Pentland, A., "Face Recognition Using Eigenfaces," *Journal of Cognitive Neuroscience,* 3(1), 1991.

[22] Viola, P., and M. Jones, "Robust Real-time Object Detection," *Technical Report CRL,* 2001.

[23] Walther, D. L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional Selection for Object Recognition - a Gentle Way," *Biologically Motivated Computer Vision - Lecture Notes in Compter Science,* Springer (2002), 2525, 472-479.

[24] Weber, M., M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition," *ECCV,* 2000.

[25] Wiskott, L., and C. von der Malsburg, "Recognizing Faces by Dynamic Link Matching," *Proceedings of the ICANN,* 1995.

[26] Wiskott, L., and J.M. Fellous, N. Kruger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. PAMI,* 19(7):775–779, 1997.