Unsupervised Fixations for Self-Organizing Neural Networks

Ryan Burt, Jose C. Principe
Computational NeuroEngineering Laboratory
Department of Electrical and Computer Engineering
University of Florida
Gainesville, Florida 32601

***Abstract*—Deep learning architectures are an extremely powerful tool for recognizing and classifying images. However, they produce the best results when trained on up to millions of tightly cropped labeled training images. To help mitigate these issues, we propose a fusion of saliency and unsupervised learning techniques, which will help to focus on relevant data and learn important features that can later be fined tuned for a specific task. In addition, by focusing only on relevant portions of the data, the training speed can be greatly improved. We test these techniques on the Cluttered MNIST database as well as the SVHN database and show how together they learn useful features in less time. The results are comparable to fully supervised methods, and can be applied to both supervised and unsupervised neural networks.**

Index Terms **- Unsupervised Learning, Saliency, Image Processing**

I. Introduction

Neural networks and deep learning architectures are the current state-of-the-art for image classification and recognition. They have been shown to reliably distinguish between as many as 1000 different classes of objects [1]. These networks, however, currently fall well short of human capabilities in two areas: recognizing objects based on a relatively small number of examples and localizing and detecting multiple objects in a single scene. In order to move towards more fully autonomous systems, we need an architecture that can extract features from a wide range of objects in cluttered scenes with minimal labels in training.

Humans have the remarkable ability to view a scene and form an overall representation in a short length of time. However, due to the complexity of visual search, it is reasonable to assume that humans do not process an entire scene at once, or even fixate on and process every small region in an image. Instead, the vision system is broken into two distinct pathways: one for spatial perception and one for object recognition. The ventral stream, or the ”what” pathway, consists of V1, V2, V4, and continues through to the inferior temporal cortex and is responsible for forming the representations and identifying objects. The dorsal stream, or the ”where” pathway, goes through V1, V2, the dorsomedial area and then the posterior parietal cortex is associated with location of objects and controlling the saccades [2].

We propose to build a two-stream perception system that is analogous to the human vision system. It will have one stream that processes an entire scene to find potentially stimulating data, and a separate stream dedicated to processing the fixations given by the first stream. By using this divide-and-conquer approach, we can quickly process large, unwieldy images and break them into smaller pieces that require the more intense computation required to extract features from an image.

The first processing pathway will replicate the dorsal stream, which deals with spatial attention over an entire scene. In the HVS, the eyes provide full access to high resolution data only in a small region called the fovea where the focus of attention is centered, approximately 3 degrees of visual angle around the point at which gaze is directed during a given moment in time. Thus, the human brain must remember and infer parafoveal and peripheral information, or use a combination of the two, to compute targets of interest for future fixation locations. As shown from empirical research on saccadic exploratory eye movements, these future fixations will target the regions in the visual periphery.

Current techniques in image processing tend to process entire images by convolving them with learned filters. By preprocessing visual data with an attention mechanism, we can focus processing only on the subregions that contain interesting data and use these to form an overall representation of the scene. In the same way that convolving learned filters over an image is a step beyond scanning pixel-by-pixel, processing still images as videos of small frames composed of visually interesting regions could be a further step that simply discards large regions of the image that have little to no effect on classification.

However, this introduces a new problem: finding the regions that contain the relevant information. Recently, methods have been proposed that suggest region both within the structure of the network [3] and as a separate mechanism based on image features [4]. These approaches are supervised and are trained to choose regions that contain data most relevant to the label.

Since the introduction of Itti's method in 1998 [5], saliency has become a popular way to predict visual attention in images and could therefore be used to segment out the interesting regions for faster processing. Saliency is defined as the state or quality by which an object stands out relative to its neighbors. An object tends to be more salient if it is brightly colored, flashy, and altogether different from its surroundings. By using saliency as a proxy for visual attention, it could be possible to create an unsupervised system that quickly selects regions of interest for more computationally intensive processing, then combine these representations into an overall understanding of a complex scene in much the same way the HVS works.

With saliency functioning as the attention system, the second pathway will replicate the ventral stream and will form representations of and extract features from objects. The ventral stream receives visual data from the fovea and builds a representation through the visual cortex that is then sent to the working memory of the scene. This is the role that neural networks and deep learning have traditionally played in image processing. By focusing the representation on only specific objects rather than the entire scene, we save computation on background and other confusing data that could be other, separate objects. By segmenting objects around highly salient points found in the attention mechanism, we can restrict the role of the network to finding invariant representations of the objects that it encounters.

These neural networks are generally trained on large datasets such as Imagenet [1] or MNIST [6] that contain tens of thousands up to millions of labeled images.By backpropagating the errors in the class labels through the network, the network is able to learn to extract the relevant features for predicting the label associated with the image. However, this learning becomes harder when multiple objects are contained within each image, each with its own label. In addition, supervised training requires labels for each image, which requires curating these large datasets and hampers their ability to be implemented outside of certain situations.

Recently, there has been study on forms of supervision other than class labels, which often consist of large datasets curated by hand. Temporal supervision [7] [8], egomotion [9], and other self-supervised or un-supervised learning techniques can be used to extract features from data before fine-tuning a network for a specific task based on a much smaller labeled set. By reducing the number of labels needed to still produce acceptable results, they have moved the networks one step closer to wider implementation in a range of problems without nicely labeled sets.

One such alternative approach is to using supervised networks that require large amounts of labeled data for training is to include temporal information. Chalasani and Principe introduced a framework that uses inference in time between subsequent frames of a video to learn relevant features in an unsupervised manner [10]. Using this framework, it is possible to learn to extract features from unlabeled datasets by taking advantage of the structure introduced through video.

In this paper we propose a new unsupervised dual pathway architecture for vision systems that separates the spatial perception element from the object recognition. The attention system will be based on a saliency measure, while the feature extraction will come from temporal supervision through the DPCN. Section II discusses related work, Section III outlines the methods used for the full vision system, Section IV presents the results, and Section V concludes the paper.

II. Related Work

A. Attention Systems

Despite these recent advances in image processing, classification results on image datasets with multiple objects in complex scenes has stagnated when using only these convolutional methods. Recently, research has begun into breaking images down into regions, then performing classification on these rather than the entire image [11]. By fusing these region detection algorithms with the recent advances in convolutional networks, classification performance on datasets such as VOC2012 have improved by up to 30% [4].

Most of the region classification methods proposed at this time were designed to be trained in conjunction with deep convolution networks, such as OverFeat [12]. OverFeat consists of a single convolutional network that is applied at multiple locations via a sliding window before producing a distribution that predicts the bounding box containing the targeted object. Alternatively, the R-CNN uses a separate region proposal method (selective search), before separately sending these regions to a CNN for classification and then finally recombining similar regions [4]. Despite the different paradigms, processing smaller regions of images has the potential to be the next breakthrough in computer vision by reducing the brute force sliding windows in the CNNs. The Spatial Transformer Network, on the other hand, integrates a differentiable image transform into the overall network structure that is capable of learning which features in an image best discriminate objects by their labels, focusing in on these objects accordingly [3].

Saliency is often used as a predictor of bottom up attention. Most saliency measures work by combining a number of simple features such as color, intensity, and orientation to find distinct regions in images that could attract the human eye. Three competing views of saliency are the center-surround methods that compare a local center to a neighborhood [5], [13], [14], [15] [16], the global context methods that compare regions to other regions from any location in the image [17], [18], and the normal image methods that compare an image to a standard ideal [19], [20], [21], [22], [23], [24].

Saliency metrics have been used in an effort to reduce computation in image and video processing, often in lossy compression algorithms that keep high resolution data only in salient areas [25] [26].

B. Feature Extraction

Training deep learning architectures without explicit class labels has been a growing area of research [8] [27]. In an effort to expand these techniques beyond datasets that come with an excess of labeled examples, there have been effort toward learning features based on other forms of supervision such as temporal and egomotion.

Goroshin et. al [7] and Wang and Gupta [8] learned short term dependencies between subsequent frames in video. Agrawal et al. modeled the egomotion of the camera in order to provide a form of supervision other than labels [9].

The DPCN by Chalasani and Principe [10] [28] used temporal predictions to learn features through time and build representations of video streams. This work was later extended into the RCPN, recurrent convolutional predictive networks, which use a dual-stream autoencoder structure to represent the current frame and predict the next frame.

III. Methods

Humans experience even static scenes through movement, whether by moving fixations across a painting or walking around a still landscape. This motion is inherent to understanding our environment; despite the lack of change in the physical properties of the scene, the information sent to the visual cortex through the eyes is constantly changing at a slow pace as the viewpoint is updated. The temporal coherence builds the full understanding of the scene as objects are recognized and placed into memory as the brain searches out new fixations.

To mimic the dual pathways of the HVS, we propose to use two separate systems: one for attention and detection of visually stimulating regions, and one for representing these regions and extracting useful features. Additionally, to avoid the use of class labeling and to leverage the capabilities gained by comibining these two systems, we will use an unsupervised saliency based detection system with a temporeally supervised learning structure.

A. Gamma Saliency

An effective attention mechanism in a dual pathway vision system should meet a few basic requirements. First, the calculation should be down quickly so that the attention works to speed scene recognition, not slow it by compounding the data. Second, the system should function as an accessory to the recognition system, not consist of one itself. This means that the attention will not be driven by recognizing objects and then assigning saliency scores. Since the dorsal stream of the HVS uses the peripheral, and therefore blurred, vision as the input to determine fixations, the system should be able to work with only low level features.

To accomplish this, we will use a simple center surround saliency method that computes local differences in regions at different scales. Although high level saliency methods exist which predict human fixations very well, these often require extensive training, require full object recognition, and are slower to compute than the more classic

Outside of saliency, gamma kernels have been used for target detection [29]. The circular shape of the gamma kernels is ideal for comparing a center region to a local neighborhood, and the size of each can easily be controlled through the use of two parameters, which allows for easily changed scales. In addition, the gamma kernel has many properties such as the ability to be computed recursively and the smoothness of the neighborhood that make it well suited to signal processing methods.

Similar to the Itti method and others, Gamma saliency is based on the center surround principle: a region is salient if it is different from the surrounding neighborhood. In order to compute these local differences, we use a 2D gamma kernel that emphasizes a center while contrasting it with a local neighborhood through convolution:

$$g_{k,\mu}(n_1,n_2) = \frac{\mu^{k+1}}{2\pi k!}\sqrt{n_1^2+n_2^2}^{\,k-1} e^{-\mu\sqrt{n_1^2+n_2^2}} \quad (1)$$

For this kernel, n_1 and n_2 are the local support grid, μ is the shape parameter, and k is the kernel order. Using μ and k, we can control the shape of the kernel: when $k = 1$ the kernel peak is centered around zero. For larger kernel orders, the peak is centered k/μ away from the center. In addition, smaller values of μ will increase the bandwidth of the peak.

With these parameters we can construct a 2D shape that compares a center region to a surrounding neighborhood by subtracting a kernel with order $k > 1$ from a kernel with order $k = 1$. The 1st order kernel functions as the center while the higher order kernel forms the surrounding neighborhood. By adjusting the shape parameter and order of the neighborhood kernel we can control the size and location of the neighborhood relative to the center, and as well as adjust the size of the center by using the shape parameter for the center kernel. Fig. 1 shows an example of a center kernel with parameters $\mu = 1$ and $k = 1$ along with the surround kernel with parameters $\mu = 1$ and $k = 10$.

For a multiscale saliency measure, we simply combine multiple kernels of different sizes before the convolution stage (2). A kernel with a larger center scale is subtracted by a surround kernel with a larger and further removed neighborhood, effectively searching for larger objects by comparing more overall area in the image. By summing all the kernels before the convolution stage, we create a system which is capable of computing saliency at different scales without adding extra computation beyond a simple summation. The kernel summation is described in (2), where all k for even m are 1 to create the center kernels. The number of different scales is one half times the length of m.

$$g_{total} = \sum_{m=0}^{M-1} = (-1^m) g_m(k_m, \mu_m) \quad (2)$$

In addition to the circular shape of the neighborhood, the gamma kernel has other useful properties that can be exploited. The shape of the neighborhoods is smooth, which is in contrast to other methods which choose neighborhoods that sample at a fixed radius. Also, the gamma kernel can be computed recursively. Though we don't make use of the recursive computation here in favor of pre-computing the kernel for speed, the recursive property could be exploited to extend this method to work in a temporal structure such as video saliency.

With this local difference measure, the rest of the saliency measure is constructed similarly to the other center surround methods [30]: the image is broken into feature matrices, each matrix is convolved with the multiscale kernel, the matrices are combined and exponentiated to accentuate peaks, then postprocessing is performed to boost results using a Gaussian blur and a center bias.

The feature matrices are composed of the CIELab color space, which has three matrices - one luminence matrix and two color opponency matrices.

In CIELab space, the distance between two colors can be calculated using the Euclidean distance, which is a useful property that we take advantage of in the convolution. Each of these matrices is convolved with the multiscale gamma kernel to get the saliency measure in each channel (3). In the following equations, • is the convolution operator.

$$S = \frac{|g \bullet L|^{\alpha} + |g \bullet a|^{\alpha} + |g \bullet b|^{\alpha}}{3} \tag{3}$$

Once we have the overall combined saliency map, there are a few common postprocessing mechanisms used to improve results. First, the main peaks in the measure are accentuated by raising the combined map to a power $\alpha > 1$. Next, it is well known that humans tend to fixate on the center of images, so a Gaussian weighting is applied to the center of the image where the variance of the Gaussian is dependent on the image size. Finally, to reduce the effects of noise and created a more streamlined map, the map is blurred using a small Gaussian kernel (4) as in [31].

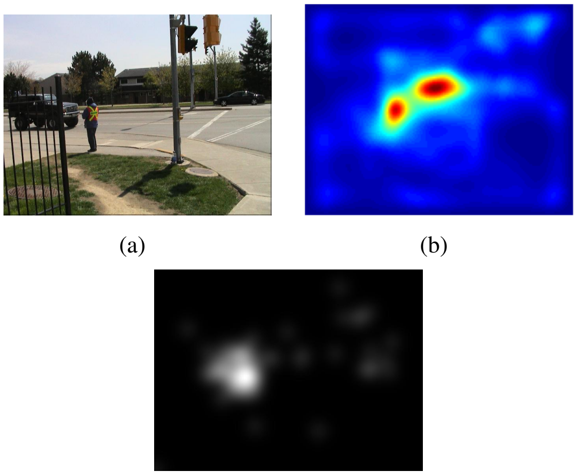
$$S = (S * G(\sigma^2)) \bullet G(.5) \tag{4}$$

B. RCPN

Rather than using explicit labels in the form of class supervision, our representer will use architectural constraints along with the structure inherent in a video stream in order to extract robust features from images. To do this, we will use the DPCN, which uses a combination of a stateless convolutional autoencoder and a convolutional RNN that encodes a dynamic state that describes the change between two frames. By using the same decoder at the end of each stream, the representations are forced to project to the same space and the error can be minimized. The cost function for the DPCN is given by

$$L_t = E[(x_{t1}D(E(x_{t1})))^2 + (x_t D(R(x_{t1})))^2], \tag{5}$$

where x_t is the video stream, the stateless encoder is E, the shared decoder is D, the CRNN by R, and E denotes the expectation operator. The architecture is trained using backpropagation.



(a)

(b)

(c)

Fig. 1. Example image from the Toronto Saliency Dataset (A), saliency map produced by Gamma Saliency (B) and the ground truth fixation map (C).

C. Vision System

Gamma Saliency and the RCPN will together form the core of the vision system, similar to the dorsal and ventral pathways in the brain, respectively. However, Gamma Saliency works on still images while the RCPN uses both spatial and temporal context to form representations of images. Therefore, to create a temporal context from which the RCPN can learn, the output of the spatial attention mechanism must be a sort of structured video from which the RCPN can infer the transitions.

In order to successfully take advantage of the temporal benefits of the RCPN, the attention mechanism must provide not only a salient point on which to focus, but a structured series of frames encompassing the object. By showing the RCPN frames with each giving a slightly altered view of the subject, the RCPN is able to learn representations that persist across the frames. This leads to a more invariant set of features learned by the vision system that can later be used for classification or other tasks.

The procedure for creating these videos is outlined in Figure 2. Given an input image, the fixation is predicted via the most salient point. The object around this point is segmented using information from the attention mechanism. Since Gamma Saliency is a multi-scale measure, the underlying feature maps contain information on which scale the object was different from its surrounding. Working from this base, we crop a patch around the object,

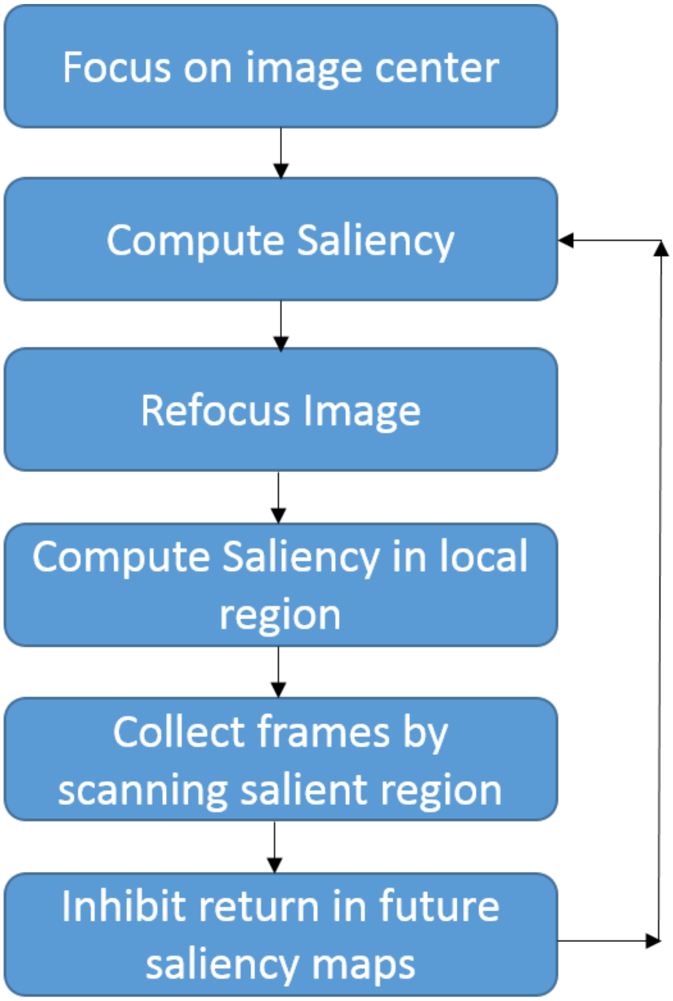

Focus on image center
Compute Saliency
Refocus Image
Compute Saliency in local region
Collect frames by scanning salient region
Inhibit return in future saliency maps

Fig. 2. Flowchart showing the focus of attention algorithm. (Needs to be cleaned)

create the video according to the data, and send that to the RCPN. There are multiple techniques that could be used for creating the videos in a structured manner. Two of these (rotating the patch around the objects and translating the frame) have been shown to be successful in previous applications of the RCPN.

Once the video of the object has been created, the fixation is moved to the next most salient point, the next object viewed and sampled, and the process continued. At each previous fixation, the saliency map is inhibited by applying an inverted gaussian that corresponds to the foveal area and the inhibition of fixation return found in the HVS.

IV. Results

A. Cluttered MNIST

For an intial test we use the cluttered MNIST dataset. This consists of handwritten digits in a 28x28 bounding box dropped into a 60x60 canvas with six pieces of correlated clutter, which are randomly sampled 6x6 frames from other digits. This functions as a very basic test of the dual-pathway vision system; the attention system should ignore the empty background and clutter to focus in only on the relevant data. This should in turn speed computation by eliminating computation over large portions of the images and help improve final classification results by helping the network focus on learning only relevant features.

This second portion could prove to be a great benefit for the self-supervised learning, since the training will not include explicit labels that the network can use to inform which features are best for classification. By using the attention mechanism to filter some of these out, we hope to create a system that learns a more robust and useful set of features than one that would try to explain the entire scene, which often contains a large amount of useless data.

TABLE1 shows the results from this test. For both supervised and self-supervised learning architectures, incoporating an attention mechanism greatly speeds the training time of the networks.

B. Multi-digit MNIST

C. SVHN

The Street View House Numbers dataset offers a tougher localization and classification challenge. It consists of over 73,000 training digits and over 23,000 testing digits in images from Google Street View. There are two main formats to the database - one cropped into 32x32 MNIST like digits with the additions of color, variable contrast, and some confusing data and the full images which contain extensive backgrouds and multiple digits in addition to the challenges in the cropped format.

Most results reported on this dataset uses the cropped digits, and even ones that try to classify the full address at once use an enlarged bounding box instead of the full image. In this test, however, we use the full images with no additional data about bounding boxes or the number of digits contained

TABLE I
MNIST RESULTS

Method \ Metric	RCPN Full	RCPN Mid)	RCPN Small	CNN Full	CNN Mid	CNN Small	STN Full	STN Mid
Classification	62.58	84.34	88.59	88.00	91.75	92.76	95.69	94.04
Time	5214	2460	1600	38	28	23	49	38

within the image. This means that our attention mechanism must localize the address, segment, then identify each digit for a success. This is a much harder problem than simply classifying boxed digits since it has combines the problems of localization and classification in a paradigm that has a range instead of a fixed output size.

V. CONCLUSION

In this paper we propose an architecture that mimics the dual pathways of the human vision system - a saliency based technique for the spatial awareness of the dorsal pathway and a network based representer for the object identification of the ventral pathway. By separating these pathways, we can achieve greater computational efficiency by quickly selecting subregions of the image for full processing, as well as improve final classification results by eliminating non-discriminatory data.

In addition, we show that unsupervised methods for each pathway compare favorably to state-of-the-art supervised methods.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[2] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[3] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 580–587.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[6] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.

[7] R. Goroshin, J. Bruna, J. Tompson, D. Eigen, and Y. LeCun, "Unsupervised learning of spatiotemporally coherent metrics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4086–4093.

[8] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2794–2802.

[9] P. Agrawal, J. Carreira, and J. Malik, "Learning to see by moving," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 37–45.

[10] R. Chalasani and J. C. Principe, "Context dependent encoding using convolutional dynamic networks," 2014.

[11] C. Gu, J. J. Lim, P. Arbeláez, and J. Malik, "Recognition using regions," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1030–1037.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[13] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[14] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, "Incremental sparse saliency detection," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 3093–3096.

[15] H. J. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*. IEEE, 2009, pp. 45–52.

[16] R. Burt, E. Santana, J. C. Principe, N. Thigpen, and A. Keil, "Predicting visual attention using gamma kernels," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1606–1610.

[17] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency based on decorrelation and distinctiveness of local responses," in *Computer Analysis of Images and Patterns*. Springer, 2009, pp. 261–268.

[18] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 1915–1926, 2012.

[19] B. Schauerte and G. A. Fink, "Focusing computational visual attention in multi-modal human-robot interaction," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM, 2010, p. 6.

[20] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[21] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "Sun: Top-down saliency using natural statistics," *Visual Cognition*, vol. 17, no. 6-7, pp. 979–1003, 2009.

[22] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, p. 32, 2008.

[23] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*. IEEE, 2009, pp. 1597–1604.

TABLE II
SVHN Results on the Full Image Dataset

Method \ Metric	RCPN Full	RCPN FOA	CNN Full	CNN FOA	STN Full
Classification	0	0	68.15	73.53	18.23
Segmentation	NA	76.93	NA	76.93	23.82
Time	0	0	2397	1506	1238

TABLE III
SVHN Results on the Bounded Dataset

Method \ Metric	RCPN Full	RCPN FOA	CNN Full	CNN FOA	STN Full
Classification	0	0	94.47	96.06	96.12
Segmentation	76.92	83.67	76.92	83.67	NA
Time	0	0	1426	1195	NA

[24] J. Li, M. D. Levine, X. An, and H. He, "Saliency detection based on frequency and spatial domain analysis," 2011.

[25] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 185–198, 2010.

[26] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on*, vol. 13, no. 10, pp. 1304–1318, 2004.

[27] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 609–616.

[28] J. C. Principe and R. Chalasani, "Cognitive architectures for sensory processing," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 514–525, 2014.

[29] M. Kim, J. W. Fisher III, and J. C. Principe, "New cfar stencil for target detections in synthetic aperture radar imagery," in *Aerospace/Defense Sensing and Controls*. International Society for Optics and Photonics, 1996, pp. 432–442.

[30] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[31] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Image Analysis*. Springer, 2011, pp. 666–675.