International Journal of Computer Vision 73(2), 159–177, 2007

DOI: 10.1007/s11263-006-9784-6

A Behavioral Analysis of Computational Models of Visual Attention

FREDERICK SHIC AND BRIAN SCASSELLATI
Department of Computer Science, Yale University, New Haven, CT 06520, USA

Received November 1, 2005; Revised June 7, 2006; Accepted July 5, 2006

First online version published in September, 2006

Abstract. Robots often incorporate computational models of visual attention to streamline processing. Even though the number of visual attention systems employed on robots has increased dramatically in recent years, the evaluation of these systems has remained primarily qualitative and subjective. We introduce quantitative methods for evaluating computational models of visual attention by direct comparison with gaze trajectories acquired from humans. In particular, we focus on the need for metrics based not on distances within the image plane, but that instead operate at the level of underlying features. We present a framework, based on dimensionality-reduction over the features of human gaze trajectories, that can simultaneously be used for both optimizing a particular computational model of visual attention and for evaluating its performance in terms of similarity to human behavior. We use this framework to evaluate the Itti et al. (1998) model of visual attention, a computational model that serves as the basis for many robotic visual attention systems.

Keywords: computational attention, robot attention, visual attention model, behavioral analysis, eye-tracking, human validation, saliency map, dimensionality reduction, gaze metric, classification strategy

1. Introduction

1.1. Motivation

Robots that interact with humans are becoming increasingly prevalent. For instance, we have robots that lead tours in museums (Burgard et al., 1998), robots that act as receptionists (Gockley et al., 2005), and robots that function as pets (Fujita, 2001). As robots begin to occupy roles traditionally reserved for people, the need for robots that are able to interact in a complex manner with human beings has increased dramatically. However, as noted by Fong et al. (2003), robots that strive to emulate meaningful relationships with humans require the ability to perceive their environments in a manner consistent with the ways humans perceive the world (Fig. 1).

For this reason, robots that seek to interact with human subjects in a general fashion often employ generalized computational vision systems based on biological inspirations (e.g. Breazeal and Scassellati, 1999). A key component of many of these vision systems is a computational model of visual attention. These computational models are, by necessity, crude approximations to the human visual attention system and typically operate by identifying, within an incoming visual stream, spatial points of interest. This computational formulation of visual attention is very limiting, in terms of the capabilities and complexities of the biological reality, as many models of visual attention could alternatively be viewed as models of eye fixation or gaze shifting. However, this restricted definition reflects the practical and operational conditions under which robots and generalized computer vision systems are found. These models serve to (i) reduce the scene to several points of particular interest, thus controlling the combinatorial explosion that results from the consideration of all

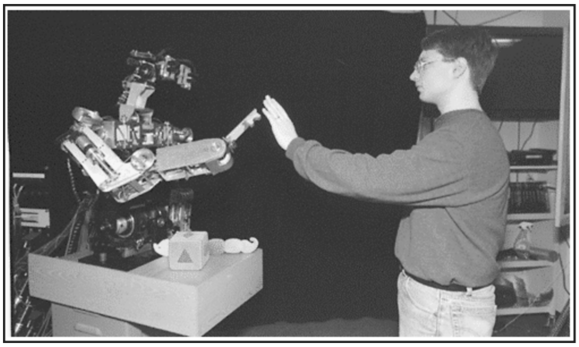

Figure 1. Robot engaging in "the Imitation Game" with a Human. A robot requires a complex visual system in order to emulate the behavior of a human subject (Scassellati, 1999).

possible image relationships (Tsotsos, 1988), and to (ii) emulate the scan-path behavior of human subjects, thus providing a naturalistic interface to behaviors such as joint attention (Scassellati, 1999; Nagai et al., 2002) and non-verbal communication (Imai et al., 2002).

However, despite the recent explosion of computational models of visual attention in recent years (half of all papers on such models have been written in the past-five years), there exists very little work regarding their evaluation. One of the difficulties of evaluating computational models of visual attention lies in the fact that, since "interesting" points are determined in a task-dependent fashion, it is difficult for a generalized model of visual attention to specify, *a priori*, what areas in a visual scene should be selected for evaluation. Three possibilities arise: (i) we could assemble a collection of scenes associated with various tasks and evaluate performance over them. This is what is done in highly-focused research areas such as face and gesture analysis, where it is clear that the "eyes" or the "hands" specify the location to which attention should be directed. (ii) By making a series of assumptions regarding the form and interactions inherent to neurobiological visual processes, we could create a generalized notion of what it means to for a particular spatiotemporal location in a scene to be "salient". This is what is done in computational models of visual attention that invoke the use of a "saliency map" (Koch and Ullman, 1985), which is a structure that assigns to each point within that visual stream some measure of visual prominence as a function of elementary modalities (e.g. color, spatial orientation, luminance). (iii) The third possibility is that we could compare the performance of a computational attention system to the computational attention system of humans.

The first possibility, creating specific measures for comparison and performance evaluation based upon expectations in specialized tasks is, of course, appropriate only for specialized tasks. In more general contexts, for example the dynamic environments required for meaningful social interactions between machines and human subjects, the formulation of a database enumerating all possible tasks would be a monstrous, if not impossible, undertaking. For this reason, specialized tests do not serve well as a basis for evaluating the visual attention systems of robots fulfilling some generalized role. The second possibility, building some coarse approximation to the biological visual attention circuitry of the human brain, is a popular approach. However, the use of a biological model for use in validation must itself be validated for biological relevance. This leaves us with comparing computational models of visual attention with human subjects.

A question arises, however, when comparing the visual scan patterns of these computational models to those of human subjects: how human is the generated eye trajectory in reality? While a particular model may generate patterns that qualitatively *appear* human-like, and subjectively *seem* to function as a realistic emulation of human objectives as reflected in the scan trajectory, rigorous empirical methods for measuring a computational model's similarity to human performance have been lacking, especially for models that perform in dynamic, naturalistic visual environments. It is our goal to define such measures of performance in order to allow the rigorous comparison of visual attention systems against both human subjects and other computational models.

In this paper we define a general framework for computational models of visual attention. This model naturally leads us to evaluative strategies for comparing computational models and human subjects. Though our framework and methods are applicable generally, as a demonstration we apply these evaluative strategies specifically to the model of Itti et al. (1998), one of the most popular computational models serving as a basis for robotic implementations of visual attention. It is hoped that our comparison methods can be used not only to help evaluate robotic attentional systems, but also to fine-tune them, as we work towards robots that interact with humans by behaving like humans.

1.2. Previous Work

In regards to biological and theoretical models of attention a large body of work exists (for specific issues relevant to this work, see Itti et al., 2005a). In regards specifically to the model of Itti et al. (1998) (which we will, from this point on, term, hopefully without offense, the "Itti model", despite the model having roots at least as far back as Niebur et al. (1995)), several quantitative analysis have been accomplished. Parkhurst et al. (2002) show that the saliency maps of images, as computed by the Itti model, is higher in locations fixated upon by human subjects than would have been expected by chance alone. Ouerhani et al. (2004) show that the saliency maps generated by the same computational attention model are correlated to approximate probability density maps of humans. In Itti et al. (2003) temporal flicker and a Reichardt model for motion are added to the Itti model, allowing for analysis of dynamic scenes. Using this augmented set of features, Itti (2005b) shows that, in short movie clips, the salience of this augmented model is higher at the target of human saccades (rapid, abrupt eye motions that shift the foveal focus from one spatial location to another) and that the motion and temporal components of the model are the strongest predictors of these saccades. Most recently, Carmi and Itti (2006) show that shortly after jump-cuts, when bottom-up influences are presumably strongest, these dynamic components have even greater ties to human saccades.

The Itti model and the interpretations of its results are not uncontroversial. Turano et al. (2003) shows that the gaze locations predicted by the static Itti et al. (1998) model are no better than random, in direct contrast to the Parkhurst et al. (2002) results. This experiment, however, uses a different measure of performance, comparing the unique model predicted gaze location to the human gaze locations, and also takes a static model and applies it to a dynamic environment. Tatler et al. (2005) employ an alternative set of elementary features as well as a different set of measures for performance to provide an alternative interpretation of the results of Parkhurst et al. (2002). Draper and Lionelle (2005) show that the iLab Neuromorphic Vision Toolkit (iLab, 2006), an implementation of the Itti model, is not scale or rotation invariant, thus questioning the appropriateness of using the Itti model as the basis of computational object recognition systems. Finally, Henderson et al. (in press) show that the Itti model can not account for human behavior during search tasks.

Though there are similarities between our study and the aforementioned work, noticeable differences exist. First, our work employs a new metric for measuring the distance between the gaze patterns of models and individuals based on classification performance and dimensionality reduction. This contrasts with studies which use Euclidean-based measures and is more similar, but not equivalent to, those studies that employ similarity based measures. Second, our work is not compatible with previous works which operate over static images (Ouerhani et al., 2004; Parkhurst et al., 2002 and subsequent discussions). The addition of a temporal component complicates analysis: human scan trajectories cannot be collapsed across the time dimension when the underlying substrate of attention, the visual scene, is time-varying. We will return to this issue in Section 2.2 and Section 3. Third, most studies choose default "mixing parameters" for the contribution of, say, color over intensity, in the final calculation of the salience map. In reality, the actual contribution of different modalities is likely to be neither strictly linear nor strictly equivalent. Computational models of attention can benefit from some optimization of parameters to match human gaze patterns, thus revealing statistics regarding the capacity of a model versus its default performance. In our work, optimization occurs as a byproduct of viewing gaze selection as a classification and dimensionality reduction problem, as we will see in Section 3.3.

We should note that, despite our focus on the Itti model, there exist many alternative computational models of visual attention both with and without motion including the work of Tsotsos et al. (1995), Wolfe and Gancarz (1996), Breazeal and Scassellati (1999), Balkenius et al. (2004), and Tsotsos (2005). The analysis of these models is not addressed in this paper due to space considerations. However, it is the ability to compare multiple such computational models that is one of the capabilities of the framework presented here.

2. Computational Models of Visual Attention

2.1. A Framework for Visual Attention

Computational models of visual attention take as an input some representation of the visual field, perform some processing internally, and return as an output a location upon which attention should be focused (Fig. 2). Typically, the internal processing can be broken up into two broad components: feature extraction and gaze computation. Feature extraction consists in

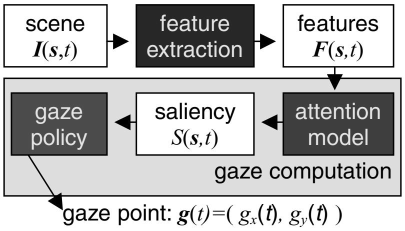
scene
$\boldsymbol{I}(\boldsymbol{s},t)$
feature extraction
features
$\boldsymbol{F}(\boldsymbol{s},t)$
gaze policy
saliency
$S(\boldsymbol{s},t)$
attention model
gaze computation
gaze point: $\boldsymbol{g}(t)=(g_x(t), g_y(t))$

Figure 2. Computational Model of Visual Attention. The spatiotemporal scene $I(s, t)$, which is a function of a spatial coordinate s and a temporal coordinate t, is operated upon by feature extraction to provide the features $F(s, t)$. The gaze computation module then takes these features and computes a gaze point $g(t)$ representing the point in the spatio-temporal scene that is most likely to be fixated upon. Gaze computation is typically broken up into two phases, an attention model that transforms features into a saliency map, and a gaze policy, that operates over the saliency map to determine the actual fixation point.

forming some abstract representation of the raw incoming visual stream and can be arbitrarily complex, ranging from simple filtering methods to systems that employ a wide range of interactions to model the pathways of the human visual system. Gaze computation consists in using the abstract representations generated by feature extraction to determine the location to which attention should be drawn. In many cases, gaze computation can be further broken up into an attention model and a gaze policy. The attention model converts the features generated by feature extraction into an intermediate representation. Often, this intermediate stage is represented as a saliency map that is proportional to, for every spatiotemporal point in the scene, the likelihood that that point will be fixated. A control strategy, the gaze policy, is then applied to the saliency map to generate a fixation point. This can be as simple as choosing the point associated with the highest salience in the saliency map. Many influential models of visual attention, such as the biologically-inspired model of Itti (1998) and the psychophysically-driven model of Wolfe (1996), as well as implementations built upon these ideas, such as the context-dependent social behavioral system of Breazeal and Scassellati (1999), obey this formulation.

We should note that, though we speak primarily of bottom-up models of visual attention (i.e. stand-alone models with predominantly forward-acting pathways from the visual scene to actual fixation), we are neither biased towards them nor limited to them in our framework. Contextual or contingent alterations to the parameters of visual attention can easily be accommodated for by this system by taking into account an augmented set of features, ones that perhaps are not associated with the visual field, per se, but instead are reflections of internal mental state, cross-modal influences, or some other set of unseen parameters leading to visually contingent behavior. Top-down behavior at the elementary level of the features themselves is also possible, as illustrated by Tsotsos et al. (1995). This type of feedback-effect between attentional model and feature extraction is somewhat more difficult to reconcile with the forward directed arrows in our model (Fig. 2), but can be accommodated for with some additional complexity. In the interest of clarity and brevity, we do not focus upon these top-down effects in this paper, but we will return to this matter in our experimental section (Section 5), where it becomes increasingly apparent that context-dependent top-down action does factor significantly in actual human visual attention.

2.2. Features for Dynamic Environments

Our framework does not depend on any specific choice of features; in fact, the utility of our framework depends on the fact that various choices of features may be compared. To guarantee a fair comparison, however, the features to be compared should all be intended to operate over the same type of scenes. Since our goal is to utilize computational models of attention in dynamic environments, such as social situations, we cannot be restricted to static images. Assuming that images from the visual stream are static suggests that motion is unimportant to visual salience, which is clearly incorrect.

Beyond the requirement that features should acknowledge that there exists a temporal dimension in addition to the spatial dimensions, we do not specify any definite form for features, except that there should exist a set of features associated with every spatial and temporal point of the spatiotemporal scene under analysis. We are then free to choose techniques for feature extraction. Sections 2.2.1–2.2.3 are examples of some of the feature sets we will use in our analysis. For simplicity, we will assume that there exist only two spatial dimensions, i.e. our spatiotemporal scenes are 2D-images that change in time.

2.2.1. Features—Raw Image Patches. Raw image patches (Fig. 3) are the simplest choice of features

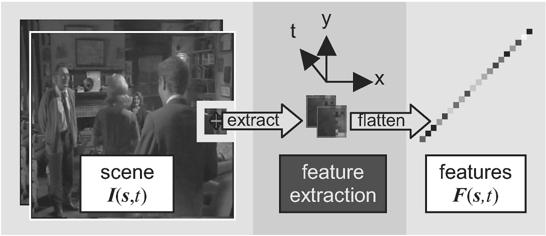

Figure 3. *Features—Raw image patches.* One of the simplest sets of features we can employ is to use all the pixels within a spatiotemporal rectangle centered at some point in the visual stream as features for that point.

associated with some particular spatiotemporal point $(\boldsymbol{s_0}, t_0)$ as:

$$\boldsymbol{F}(s_0, t_0) = \{\boldsymbol{I}(\boldsymbol{s_0} + \boldsymbol{\delta}\mathbf{s}, t_0 + \delta t)\}, \quad \forall\, \boldsymbol{\delta}\mathbf{s} \in N_s, \delta t \in N_t$$

where N_s is some set of spatial offsets, N_t is some set of temporal offsets, and the two sets together define a spatiotemporal neighborhood in the vicinity of $(\boldsymbol{s_0}, t_0)$. The use of raw image patches is inspired by the design of the eye, which has high spatial acuity at the fovea, and lower resolution towards the periphery (Tessier-Lavigne, 1991). By drawing $\boldsymbol{\delta_s}$ from some set of offsets that are tightly coupled with the immediate region surrounding a particular $\boldsymbol{s_0}$, we coarsely approximate this effect. In essence the features that draw attention to a particular point are highly connected to the history of what has transpired near that point. We choose our spatial neighborhood around a spatiotemporal point to be a square centered around the spatial aspect of that particular point, and causally from several points backwards in time: $N_s = \{(\delta_{sx}, \delta_{sy})\}\ \forall \delta_{sx} \in \{-L, L\}, \delta_{sy} \in \{-L, L\}$ for some characteristic length L.

2.2.2. Features—Gaussian Pyramid. A more satisfying alternative is to build a Gaussian pyramid of the scenes by progressive filtering (Burt and Adelson, 1983) (Fig. 4). The features corresponding to a point $(\boldsymbol{s_0}, t_0)$, then, are:

$$\boldsymbol{F}(\boldsymbol{s_0}, t_0) = \{\boldsymbol{I_i}(\boldsymbol{s_0}, t_0 + \delta t)\}, \forall i \in \boldsymbol{N}_L, \delta t \in \boldsymbol{N}_t$$
$$\boldsymbol{I_i}(\boldsymbol{s}, t) = \boldsymbol{I}(s, t) * \boldsymbol{G}^i$$

where $\boldsymbol{G}$ is a Gaussian filter, and $\boldsymbol{G}^i$ represents i convolutions of $\boldsymbol{G}$. In other words, the features at a particular point correspond to raw image information at that point, plus the image information of $n + 1$ blurred versions of the original image, $N_L = \{0 \ldots n)\}$, also at that same point. As in 2.2.1, a select set of history is retained in the temporal neighborhood N_t in order to capture the time-varying nature of the scene.

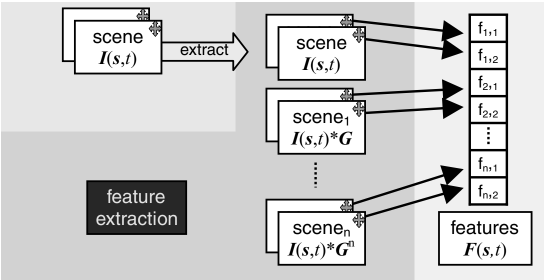

Figure 4. *Features—Gaussian pyramid.* This simple set of features associates every spatiotemporal point $(\boldsymbol{s}, t)$ with the image characteristics $\boldsymbol{I}(\boldsymbol{s}, t)$ and $n + 1$ blurred out versions of $\boldsymbol{I}$. In the above figure, $\boldsymbol{G}^i$ is convolved with itself i times. Several time points are also employed in building the features $f_{\lambda,\tau}$, where λ represents the Gaussian level and τ represents the time index.

2.2.3. Features—Biologically Inspired Models. Currently, the most popular computational models of visual attention existing on robotic platforms are based on biological inspiration. The computational models of visual attention mentioned at the end of Section 2.1 are all biologically-inspired models, differing in their choices of features, techniques for saliency computation, and, ultimately, their intended purpose. Of the models mentioned, the model of Itti et al. (1998) has received the greatest attention in application and analysis. For these reasons, we employ the Itti model exclusively in our analysis of biologically-inspired models.

2.3. The Itti Model

The Itti Model is a feed-forward bottom-up computational model of visual attention, employing, at its most basic level, decompositions into purely preattentive features. This gives advantages in both speed and transparency. It is a model that is not only simple but also rigorously and specifically defined, a strong advantage for implementation, extension, and reproducibility of results. It is also possible to download the source code for the Itti model (iLab, 2006), though we, in this study, implement the Itti model in Matlab directly from Itti et al. (1998). Note that we use the earlier version of the Itti model as a base and not the augmented version of Itti et al. (2003). We have also performed analysis over both our custom implementation and the publicly available source, a point which we return to in the discussion.

The Itti model extracts the preattentive modalities of color, intensity, and orientation from an image. These modalities are assembled into a multiscale representation using Gaussian and Laplacian pyramids. Within each modality, center-surround operators are applied in order to generate multiscale feature maps. An approximation to lateral inhibition is then employed to transform these multiscale feature maps into conspicuity maps, which represent the saliency of each modality. Finally, conspicuity maps are linearly combined to determine the saliency of the scene. These operations are summarized in Fig. 5.

The original Itti Model (Itti et al., 1998) did not include a modality for motion. This was rectified by later work (Yee and Walther, 2002; Itti et al., 2003). However, there seems to be a mismatch between the theoretical concerns of the model and the implementation (especially see Itti et al., 2003). The reasons for this discrepancy are not clear. In this work, we use a different formulation for motion saliency (see Appendix A) which resulted in better empirical performance. The differences between this formulation (the addition of which will lead to the Extended Itti Model) and previous work are subtle. However, our experience with our own implementation, including use on a humanoid robot, has shown the formulation presented in Appendix A to be both reasonable and robust.

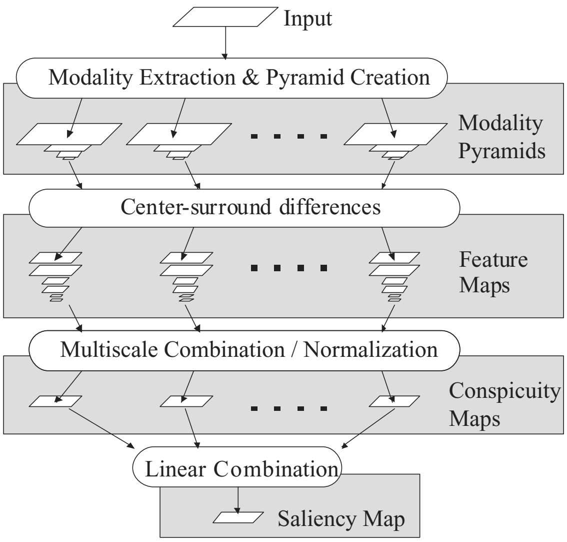
Input
Modality Extraction & Pyramid Creation
Modality Pyramids
Center-surround differences
Feature Maps
Multiscale Combination / Normalization
Conspicuity Maps
Linear Combination
Saliency Map

Figure 5. Itti Model general architecture (adapted from Itti et al. (1998)). Modalities such as color, intensity, and orientation are extracted and operated on over several stages in order to produce the saliency map associated with the input.

2.4. *The Computation of Fixation from Features*

The gaze computation process takes, as an input, extracted features, and returns, as an output, a point of fixation. As mentioned earlier, this process can be broken up into two modules: an attention model and a gaze policy.

2.4.1. Attention Model. The attention model transforms features associated with a particular spatiotemporal point into a single value that is representative of how likely that point is to be focused upon. In other words, if the original spatiotemporal scene is a color movie with three channels, and this scene is analyzed at the region level to extract D features at every point, the mapping that occurs for each spatiotemporal point is $\mathbf{R}^2 \times \mathbf{R}^+ \rightarrow \mathbf{R}^D \rightarrow \mathbf{R}^1$. The last level in this transformation is the saliency map, a notion originally formulated by Koch and Ullman (1985). We note that though there appears to be some evidence for the coding of an explicit saliency map in the brain (e.g. in the superior colliculus, Kustov and Robinson, 1996; in the lateral geniculate nucleus, Koch, 1984; in V1, Li, 2002; in V1 and V2, Lee et al., 1999; in the pulvinar, Petersen et al., 1987; Robinson and Petersen, 1992; in V4, Mazer and Gallant, 2003; in the parietal cortex Gottlieb et al., 1998; general discussion, Treue, 2003), the question of whether or not saliency maps are actually present physiologically has not been answered definitively. Here, we use the saliency map purely as a computational convenience, and where we do not denote saliency as "computational saliency", we hope that it is understood that our work primarily refers to the computational representation of saliency which may or may not have some biological correlate. Without loss of generality, however, we can employ saliency maps as an intermediate step since any computational model that generates some specific point corresponding to a point of fixation has at least one saliency map that, under some fixed gaze policy, returns the equivalent point. For example, a saliency map that is zero everywhere except at the point of fixation, where it is positive, will return the correct point under **arg max**.

2.4.2. Gaze Policy. One of the simplest gaze policies we can employ is one that simply indexes the location in the saliency map corresponding to highest peak. In other words:

$$\boldsymbol{g}(t) = \mathop{\mathbf{arg\,max}}_{s}\,(S(\boldsymbol{s}, t))$$

However, there exist many possibilities for the control strategy. For instance, if the saliency map corresponds to a probability distribution, sampling could result in non-deterministic behavior. Additional structure, such as inhibition-of-return (as is implemented in Itti et al. (1998)), can yield much more complex time-varying behavior. In any case, the ability to obtain a single point from a computational model of attention allows us to build methods for comparing gaze patterns directly.

3. Metrics for Modeling Visual Attention

To compare computational models of visual attention and human subjects, we need to define some metric over which some notion of similarity can be made. An obvious choice for such a metric is gaze fixation distance. We can say that a particular gaze process $\mathbf{G_a}$ is close to another gaze process $\mathbf{G_b}$ if the points of fixation chosen by $\mathbf{G_a}$ and $\mathbf{G_b}$ are spatially close for all points in time. However, the major problem with distance measures is highlighted in Fig. 6. Essentially, employing distance as the sole measure of similarity results in questionable results since gaze patterns are dependent on the underlying scene. Note that this is true whether we employ distance directly or use some nonlinear variant that is dependent upon distance, such as overlap of Gaussians centered at fixation points.

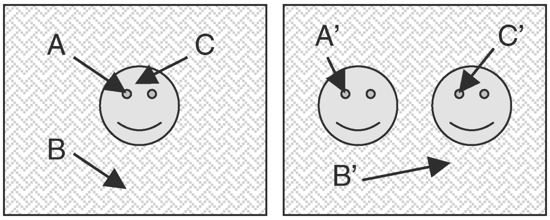

Figure 6. Problems with Fixation Distance Metrics for Measuring Similarity. In the case on the left, if some model picks point A and another model picks point B, we could safely say that these two models are dissimilar. Conversely, if one model picks A and another model picks C, we could say that the models are similar. In this case, a distance metric based on distance between fixations makes sense. In the case on the right, if one model picks A' and another model picks B', we can still say that these two models are similar. However, if one model picks A' and another model picks C', the distance is roughly equivalent to the case of A'-B'. However, the underlying features at these points are very similar. In this case, using a fixation distance metric does not make sense. By employing distance metrics between points of fixation, we ignore the underlying substrate of visual attention: that of features of the scene itself.

An alternative to using distances for comparison is to use some index of saliency as the measure. This is the method employed in both Parkhurst et al. (2002) and Ouerhani et al. (2004). Notably, both groups use the locations that human subjects fixate upon to index into the saliency map, and show that the saliency at the locations attended to by humans is greater than what would be expected by a random process. Since saliency is assembled from features, and since features change in a time-varying fashion, the technique of collapsing eye movements across time, as done by Ouerhani, is not applicable to our environment. For instance, if, on the right image of Fig. 6, we were to show only one face, and after some short time, cover that face and display the other face for a time period equal to the first face display, and if a process $\mathbf{G_a}$ focused on faces in both situations, but a process $\mathbf{G_b}$ focused on the conjugate empty space in the same situations, we would have identically collapsed probability functions, but a very different underlying gaze strategy.

Another alternative is to aggregate the looking points of a large number of human subjects for each point in time. On static images it is easy to obtain 10 seconds of looking time at 60 eye recordings a second for a total of 600 eye fixations over an image for a single individual. For time-varying images, however, we require either a large number of subjects or a set of strong assumptions about the probability fields associated with each eye fixation (such as a Gaussian region centered about each gaze fixation), to obtain the same level of sampling. Using a large sample of individuals, of course, does provide a great deal of information. However, in the interest of a generalized computational framework for visual attention, we desire a technique that, while still being able to benefit from multiple sources of data, is not completely dependent on the sampling size of human data.

Finally, for non-biologically-inspired models, saliency is not necessarily a cleanly defined concept. Since we want to compare models to models as well as models to humans, it is in our interest to develop some strategy that makes saliency somehow comparable across various selections of features. The method we employ in this work is to define distance at the feature-level. That is, we say that two spatiotemporal locations are "close" if their underlying features are close. The particular implementation of this distance measure is the subject of the next section.

4. A Classification Strategy for Computational Saliency

We desire a method for forming saliency from features (ignoring task knowledge and other top-down effects which definitely play a role in biological visual saliency, a point to which we will return later). The method that we employ in this work is to divide spatiotemporal scenes into two classes: (i) locations attended-to and (ii) locations *not* attended-to. We define saliency as some function that is related to the probability that a particular location, based solely on its associated features, is likely to be fixated upon by a human observer. By defining saliency in this manner we achieve several goals: (i) we obtain a mapping from features to saliency that corresponds to a structured and intuitive measure of distance in feature space; (ii) we obtain a method that makes saliencies for different choices of features comparable, since they represent an underlying likelihood of fixation; and (iii) since features are translated directly into saliencies, which, in turn represent, in some fashion, probabilities, we do not need to optimize an individual model to match a human's gaze pattern—such an effect is incorporated implicitly in the mapping.

4.1. Bayesian Classification Strategy for Attention

We know that, for some feature vector $\boldsymbol{f}$ and class c_i:

$$p(c_i \mid \mathbf{f}) = \frac{p(\boldsymbol{f} \mid c_i)\, p(c_i)}{p(\boldsymbol{f})}$$

If we were to use a Bayesian classifier, we would, for two classes c_0 = attended-to and c_1 = not attended-to $= \neg c_0$, choose class c_0 if $p(c_0 | \boldsymbol{f}) > \ominus\ p(c_1 \mid \boldsymbol{f})$ for some threshold θ, and would choose class c_1 otherwise. We could thus define saliency to be:

$$\varphi(\boldsymbol{f}) = \frac{p(\boldsymbol{f} \mid c_0)}{p(\boldsymbol{f} \mid \neg c_0)}$$

However, φ can be arbitrarily large, due to the term in the denominator. More problematic is that $p(\boldsymbol{f} \mid c)$ must be estimated. This tends to be quite difficult in high dimensional spaces, and, even in low dimensions, may require more complicated approximation techniques such as mixture-of-Gaussians. Note that this formulation is similar to work by Torralba (2003) and Itti and Baldi (2006a), both of whom take a Bayesian approach towards aligning scene features with points of regard.

4.2. Fisher's Linear Discrimiant Strategy

Though useful as an intuitive conceptualization of the visual attention process, it is not necessary to explicitly form a probability map representing the likelihood of attending to each spatiotemporal location. Attention is directed towards some "interesting" point. In some ways, it does not matter if the function governing the decision to attend to some location is twice or three times the value of some other, less likely to be attended-to, point, only that it be greater. For this reason we can relax the ideal that saliency should correlate directly with a probability, and use forms of dimensionality reduction to aid in the computation of salience. Many dimensionality reduction schemes exist, with varying abilities to adapt to non-linear relationships, and with varying levels of biologically plausibility. The method that we employ here is one of the oldest, and simplest, techniques: Fisher's linear discriminant.

By using this model, we do not presuppose the existence of any biological or psychophysical effect, and, furthermore, only need to specify that we expect some difference exists between the locations that are attended to and the locations that are not. With the two classes, c_0 and c_1, corresponding to points in the spatial temporal scene where gaze is fixated and points where gaze is not fixated, respectively, maximization of the Fisher criterion function:

$$J(w) = \frac{w^t S_B w}{w^t S_W w}$$

yields the solution:

$$w = S_W^{-1}(m_1 - m_2) \tag{1}$$

where

$$m_i = \frac{1}{|c_i|} \sum_{x \in C_i} x$$

and

$$S_W = S_1 + S_2$$

with

$$S_i = \sum_{x \in C_i} (x - m_i)(x - m_i)^t = (|c_i| - 1) \sum_i = k_i \sum_i \tag{2}$$

(Duda and Hart, 2001). The projection matrix w is used to project a location's features to one dimension, and it is this projection that serves to approximate saliency.

Of particular concern, however, is the fact that there is a large asymmetry in the sizes of the populations of classes. At any particular point in time, there are a large number of regions corresponding to points *not* fixated upon, but only one region corresponding to that point upon which gaze *is* fixated. If we were to take Eq. (2) verbatim, then we would end up with a projection predominantly shaped by the covariance of patches that gaze is not fixated upon, and this, in turn, would be similar to the general properties of the spatiotemporal scenes in the data set we choose. Thus the hypothesized difference in covariance structure between gaze-fixated points and non-gaze-fixated points would tend to be washed away. For this reason, we instead assume that the factors k_i in Eq. (2) are equal across classes. This assumption leads to greater discrimination ability between fixated and non-fixated points, as measured empirically.

4.3. Rank Ordering

The measure from Eq. (2), a value for saliency at every point, is a projection not in metric proportion. For example, if application of our weight matrix to some map of features were to yield a particular value at one point, and half that value at a second point, we should not interpret this to imply that the second point was half as likely to be focused upon. We could reasonably assume, however, that the second point was less likely to be focused upon. For this reason, instead of using the *value* of saliency directly, we examine our samples in terms of their *ordering* at a given point in time. In other words, we assume that, for a given time t, the saliency at any spatial location s can be compared with the saliency of other spatial locations via ranking. We do not assume that saliency computations at different points of time can be directly compared, as this would imply that our saliency measure in some way represented some global metric with global implications rather than a local metric over local features.

It is important to note that many alternative strategies exist for normalizing saliency. We could, for instance, require the saliency map span a range of values from 0 to 1, or that the energy of the saliency map be normalized. These normalization strategies require different sets of assumptions. Our choice of attention model will greatly impact these relationships. For instance, the Bayesian strategy presented in Section 4.1, formulated as a ratio, and the Fisher discriminant strategy presented in 4.2, formulated as a projection, result in two very different distributions. Since maintaining comparability despite changes in the underlying attention model or feature extraction process is one of the goals of this work, we employ a final measure that is independent of monotonic transformations on saliency.

5. Experiments

We want to compare different computational models of visual attention against human subjects. However, since we are comparing multiple models, we must also control for multiple sources of variation, such as the inherent dimensionality and spatiotemporal extent of the underlying features. By comparing a wide range of parameters on our computational models, and by choosing good controls for our human subjects, it is hoped that these sources of variation can be controlled.

5.1. Data and Subjects

The human subjects in this experiment consist of 10 individuals drawn from a population of adolescents and young adults that are intended to serve as age and verbal-IQ matched controls for a different study, one which compares these controls versus individuals with autism (Klin et al., 2002). While this group is predominantly considered normal, some of the individuals of the population fall in a range that labels them as mildly mentally retarded. It is our intent to conduct this experiment over subjects that are slightly varied in mental capability, as we do not expect our technique to hinge on a notion of a "typical" human subject.

The gaze patterns for these human subjects are obtained via a head mounted eye-tracker (ISCAN Inc, Burlington, Massachusetts) under controlled conditions as the subjects watch two different, approximately 1 minute long, clips of the 1966 black and white movie "Who's Afraid of Virgnia Woolf". The eye tracker employs dark pupil-corneal reflection video-oculography and has accuracy within $\pm 0.3^{o}$ over a horizontal and vertical range of $\pm 20^{o}$, with a sampling rate of 60 Hz. The subjects sat 63.5 cm from the 48.3 cm screen on which the movie was shown at a resolution of 640×480 pixels.

All gaze data, except for locations which were invalid due to technical or experimental issues, were used in subsequent analysis. That is, the results were generated from gaze points that were not segregated into saccades and fixations. The use of a simple velocity threshold criteria for saccade-fixation segregation (Salvucci and Goldberg, 2000) with the cut-off set to 30 degrees $\sec^{-1}$ and subsequently labeled saccades removed from study did not change our basic findings, but did improve results across the board for human subjects. Though the effect was small, this finding is consistent with the theory that visual processing does not occur during saccades. Since the use of a saccade identification scheme did not impact our results, in this work we omit consideration of saccade identification reasons of economy, with the understanding that the use of an appropriate fixation and saccade identification scheme is both relevant and important to a computational model that seeks to describe human gaze patterns.

To assess the performance of human subjects versus chance, it is necessary to define comparative data sets that are basically uncorrelated to human subjects. However, we believe that it is not sufficient to simply sample random points, or to compute statistics over the entire saliency map, to generate our control data. Our set of synthetic data consists of several different types of random gaze strategies:

(1) *random filters (RF)* – these correspond to a random weight matrix (Section 4.2). These are projections that are completely uncorrelated with any events in the visual scene.

(2) *random saccades (RS)* – these scan paths are created by an algorithm that waits at a given spatial location for some time and intermittently jumps to new locations. The decision to jump is assessed probabilistically, and the distance and angle of jump are generated from uniform distributions randomly.

(3) *random physiological (RP)* – these scan paths are created algorithmically from physiological gaze measurements using a probabilistic model. The spatial gaze location of the RP scanpath as a function of the current movie frame number t is $\boldsymbol{g}(t)$, with $\boldsymbol{g}(0) = \boldsymbol{s_0}$ where $\boldsymbol{s_0}$ is the center of the screen. At each new frame the gaze location is updated according to $\boldsymbol{g}(t+1) = \boldsymbol{g}(t) + \Delta(d(t))$, where Δ is a function that takes a step as determined by the distance traveled $d(t) = \|\boldsymbol{g}(t) - \boldsymbol{g}(t-1)\|$. Δ is spatial update in a polar frame, $\Delta = (\mathrm{d}r \cos(\mathrm{d}\alpha), \mathrm{d}r \sin(\mathrm{d}\alpha))$, where $\mathrm{d}r = d(t+1)$, and $\mathrm{d}\alpha$ is the change in angle $|\alpha(t+1) - \alpha(t)|$. $\mathrm{d}r$ is calculated by a heuristic that samples from the distribution $p(d(t+1)|d(t))$, the dependence of the current velocity on previous velocity. This incorporates the idea that when velocity is high (as in a saccade), it is more likely that movement will continue to be high, and when velocity is low (as in microsaccades during a fixation), it is more likely that movement will continue to be low. The heuristic used is a spill search followed by random sampling: we first locate all time points in the physiological samples where the distance traveled during a given frame was closest to $d(t)$ plus or minus some spill fraction (e.g. 5% of all indices, centered at $d(t)$). We then sample randomly from this collection to get $\mathrm{d}r$. Similarly, the change in angle $\mathrm{d}\alpha$ is calculated by sampling from the distribution $p(\mathrm{d}\alpha(t+1)|d(t+1), d(t))$, where joint proximity to $d(t+1)$ and $d(t)$ is calculated in the Euclidean sense. The dependence of $\mathrm{d}\alpha$ on both previous and current distance reflects the interaction between deflection and velocity in gaze patterns.

The parameters of this synthetic control data are varied in order to span a space of random behavior, and $N = 5$ synthetic sets are generated for each random gaze category.

5.2. *Methods*

The modified Itti Model with motion, which we employ in our analysis, computes its output on a specific sized image. Since we want our results to be comparable spatially, we first begin by downsampling our input stream so that all saliency maps will match the final size of the Itti model. That is, for raw pixel and Gaussian pyramid techniques, we downsample each image in the stream from 640 × 480 pixels to 40 × 30 pixels. This results in a fairly coarse spatial resolution, implying a not inconsequential degree of blurring. However, we have found, with all other parameters held constant, that this blurring increases the performance of our models, likely due to two reasons: (i) it effectively eliminates error due to the inherent inaccuracy of the eye tracking technology used, and (ii) downsampling increases the spatial span of our features. The tradeoff between the information lost and spatial range gained is an issue that we hope to address in future work.

Next we apply our various computational models of visual attention to generate features associated with every spatiotemporal point in the visual scene. For every model, the features will be drawn from two time points: $\{-100\ \text{ms}, -300\ \text{ms}\}$. In other words, the features associated with a spatiotemporal point consist of features extracted from the history of that spatiotemporal point, since gaze fixation is not an instantaneous operation, but instead occurs shortly after some salient event or feature is detected. Though we will vary the parameters of our computational models, we will adopt some standards for each model we employ:

(1) *raw image patch features* – patches are always centered on some pixel, are always square, and contain an odd number of rows and columns. Since our input stream is black-and-white, there is only one dimension associated with each spatiotemporal point: intensity. Each raw image patch is then (*length* × *width* × *point dimensions* × *temporal dimensions*) $= N \times N \times 1 \times 2 = 2N^2$ dimensions.
(2) *Gaussian pyramid features* – we will always employ 3 total levels in our pyramid. If we need to vary the dimensions of this model, we use data from the pyramids in adjacent locations, as we do for raw image patch features. Each pyramid patch is then $N \times N \times 3 \times 2 = 6N^2$ dimensions.
(3) *Extended Itti Model features* – We use modalities of intensity, orientation, and motion. We omit the color modality since the images are black and white. As with the other features, if we need to extend the dimensionality of the Extended Itti Model we use features from spatially adjacent cells. Each feature associated with some spatiotemporal point as computed by this Extended Itti Model is $N \times N \times 3 \times 2 = 6N^2$ dimensions.

After we obtain our features, we compute for human and synthetic data sets the optimal filters (as described in Section 4.2) for each individual, where an individual is represented by some gaze trajectory over the spatiotemporal scenes. Naturally, we exclude *random filters* from this process. We begin by assembling the set of attended-to features by indexing the features found at each spatiotemporal location that a particular individual's gaze is directed. Next we obtain the set of *not* attended-to features by indexing the features *not* found at the gaze locations of that particular individual (Fig. 7). We do this by randomly sampling locations that are drawn from the pool of points some minimum distance away from the attended-to location (this distance is, in this work, 31.25% of the number of rows). The minimum distance requirement helps to make the two distributions distinct, as it is known that image patches in natural images are correlated to their distance. 15 *not* attended-to locations are sampled for every attended-to location. Together these sets of features enable us to compute a filter for each individual by finding the optimal projection as given by Eq. (1).

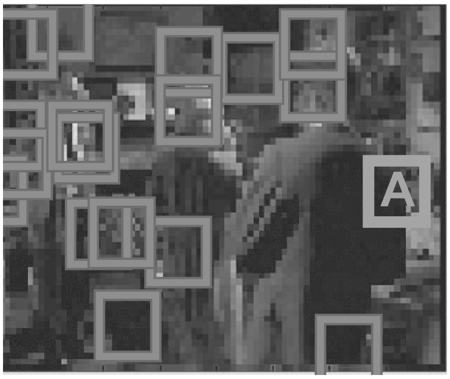

Figure 7. Extracting features for attended-to and not attended-to locations. The rightmost box (marked with A) is centered at the attended-to location. The boxes found to the left (unmarked) of the rightmost box are randomly sampled points that are used to generate the *not* attended-to pool.

By taking these filters and applying them to the underlying features found at each point in the visual scene, we can obtain saliency maps tuned to each individual as they watch some particular movie clip. As discussed in Section 4.3, we rank order the saliency maps spatially for every time point to obtain rank-ordered saliency maps, and use this as our comparative function. In other words, given an optimal weight W_u computed for some individual u's gaze pattern, we can calculate the time-varying saliency map $S_u(s,t)$ tuned to that individual u:

$$S_u(s,t) = W_u * F(s,t)$$

We then compute, for each frame in the movie, the rank percentile of v's gaze fixation on S_u. That is we find:

$$s_{u,v}(t) = S_u(g_v(t), t)$$

$$r(x, thr) = \begin{cases} 0, & x \geq thr \\ 1, & \text{otherwise} \end{cases}$$

$$R_{u,v}(t) = \frac{\sum_{i \in I} r(S_u(i,t), s_{u,v}(t))}{|I|}$$

where $g_v(t)$ is the spatial gaze location fixated upon by user v at time t, I is the set of valid spatial locations in the spatiotemporal scene, and $R_{u,v}(t)$ is the rank percentile score at time t of v's gaze fixation on the saliency map as trained by user u.

Since we are interested in comparing overall performance and group effects, we then generate a receiver operator characteristic (ROC) curve for $R_{u,v}(t)$ as a function of response percentile *rp*. That is, in order to compute the overall goodness of fit of u's model on v's data, we sort all the frames corresponding to $R_{u,v}(t)$ and sample the sorted list at various response percentiles *rp*.

Finally, this information is aggregated into groups and compared. We are interested in both individual effects as well as group effects, and can obtain these measures by utilizing the filters of one individual on other individuals (i.e. u and v in the above formulation do not have to be the same). For instance, we examine the performance of a human individual's filter on the individual himself ("matched" datasets, group *n ma* in figures and tables), as well as the performance of a human individual's filter on other individuals ("unmatched" datasets, group *n un* in figures and tables). We also examine the filters of the synthetically generated random filters, random saccades, and random physiological simulations, when applied to human individuals (*rf/n*, *rs/n*, and *rp/n*, respectively). We train on every other frame of only one particular movie. This allows us to test upon the frames not trained upon as well as on a separate movie that does not overlap temporally.

5.3. Results

By tuning our models to each individual human subject as well as all synthetic data, we are able to generate ROC curves as shown in Fig. 8. We are able to compute cross statistics over different training/testing pairs (i.e.

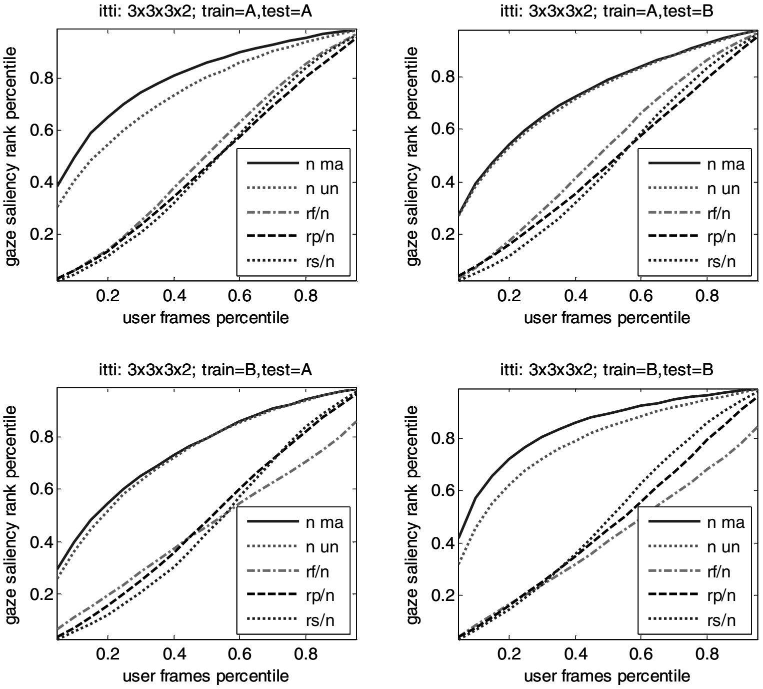

Figure 8. ROC curves for the one variation of the Extended Itti Model. This model extracts 3 modalities (orientation, intensity, and motion) over a 3 × 3 patch that extends across 2 time points. The various curves represent different training/testing variations. For instance, the upper right image are statistics for individuals trained on a movie clip A and tested on a different clip B. The y-axis in all graphs is the average saliency rank percentile reported for a particular individual's gaze trajectory. The x-axis is the percentile of all frames that fall beneath the corresponding saliency rank percentile. The categories are: n ma = human subjects trained on themselves and tested on themselves; n un = human subjects trained on themselves and tested on all other human subjects; rf/n, rs/n, rp/n = randomly generated filters, synthetic saccades, synthetic random physiological simulations, respectively, each applied to human subjects. These graphs show that the synthetic trajectories, when trained and applied to human subjects, perform basically at chance level, in comparison to human trajectories trained and tested on other human subjects. The results also show that tuning is tied to particular spatiotemporal scenes and does not transfer across data sets.

Table 1. Median gaze saliency rank percentiles for variations of computational models of visual attention. Each table represents a single testing/training pairing (e.g. A on B implies that the table represents models trained on data set A and tested on data set B). The columns L is the spatial length (in pixels), and ND is the number of dimensions associated with that particular model's features. The categories are the same as those used in Fig. 8, and the values are reported in mean percentages with standard deviations.

A on A	L	ND	*n* ma	*n* un	*rf/n*	*rs/n*	*rp/n*
Raw	1	2	74 ± 5	74 ± 5	66 ± 11	50 ± 20	54 ± 23
Raw	5	50	86 ± 3	80 ± 4	53 ± 15	50 ± 10	51 ± 18
Pyramid	1	6	82 ± 3	81 ± 3	47 ± 24	52 ± 19	53 ± 23
Pyramid	3	54	87 ± 2	80 ± 4	52 ± 23	52 ± 9	52 ± 17
Itti	1	6	79 ± 4	79 ± 3	45 ± 24	63 ± 14	58 ± 25
Itti	3	54	86 ± 2	80 ± 4	56 ± 16	44 ± 10	46 ± 14
A on B	*L*	ND	*n* ma	*n* un	*rf/n*	*rs/n*	*rp/n*
Raw	1	2	77 ± 6	77 ± 6	69 ± 12	51 ± 24	55 ± 27
Raw	5	50	86 ± 3	85 ± 4	55 ± 17	53 ± 9	46 ± 22
Pyramid	1	6	87 ± 3	87 ± 4	47 ± 27	56 ± 22	48 ± 29
Pyramid	3	54	86 ± 3	85 ± 4	48 ± 24	55 ± 9	45 ± 21
Itti	1	6	78 ± 4	77 ± 3	48 ± 22	60 ± 18	58 ± 23
Itti	3	54	78 ± 4	78 ± 5	55 ± 15	43 ± 13	46 ± 17
B on A	*L*	ND	*n* ma	*n* un	*rf/n*	*rs/n*	*rp/n*
Raw	1	2	74 ± 5	74 ± 5	45 ± 24	45 ± 22	55 ± 23
Raw	5	50	81 ± 3	80 ± 3	47 ± 17	49 ± 7	55 ± 8
Pyramid	1	6	80 ± 4	80 ± 4	46 ± 26	44 ± 19	67 ± 15
Pyramid	3	54	80 ± 3	80 ± 3	62 ± 18	47 ± 9	54 ± 9
Itti	1	6	80 ± 4	80 ± 3	66 ± 11	44 ± 17	51 ± 19
Itti	3	54	80 ± 4	79 ± 4	58 ± 22	43 ± 8	48 ± 13
B on B	*L*	ND	*n* ma	*n* un	*rf/n*	*rs/n*	*rp/n*
Raw	1	2	77 ± 6	77 ± 6	45 ± 28	44 ± 26	56 ± 27
Raw	5	50	91 ± 2	87 ± 3	51 ± 21	51 ± 11	51 ± 13
Pyramid	1	6	88 ± 3	87 ± 3	45 ± 31	43 ± 25	67 ± 21
Pyramid	3	54	91 ± 2	88 ± 4	65 ± 23	47 ± 9	49 ± 16
Itti	1	6	81 ± 5	79 ± 5	65 ± 11	47 ± 17	52 ± 18
Itti	3	54	89 ± 2	84 ± 4	54 ± 25	49 ± 11	45 ± 13

training on one movie, testing on another), as well as over different groupings of human and synthetic data (Table 1). This reveals several findings.

First, human subject tuning is better than random even for the largest reported synthetic result ($p < 0.05$). In other words, chance, or some general artifact of our processing technique, can not account for the performance of any model of visual attention that is tuned to human subjects.

Second, if we examine the matched versus unmatched human performance across models, we see that, for models trained and tested on the same movie clip, differences appear only as the number of dimensions of the models increase. This suggests that our computational models of visual attention are being tuned to general, rather than specific, strategies at low dimensions. For instance, if we look at the data with the lowest number of dimensions in Table 1, that of raw patches of length 1, we can see that matched performance is equivalent to unmatched performance for all cases. This implies that, in this case, tuning the model to a particular individual does not provide greater specificity. When we boost the dimensionality of our features to around 50, however, we see that, for models tuned to particular individuals and tested within the same data set, greater specialization is achieved.

This brings us to the third point: when we apply tuned models to gaze trajectories obtained over different data sets, all differences between matched and unmatched subjects disappear. This suggests that tuning is specific to the spatiotemporal scene over which the model is trained, and that the effects of tuning, when they are apparent, disappear as we move further from the training source (Fig. 9). At some basic level, this implies that the actual parameters of these computational models of visual attention are time-varying, suggesting that top-down or contextual effects upon visual attention are observable and significant. In Fig. 10 we can see this more clearly. When the focus of a human individual shifts from the person who is talking to the person who is being talked to, the model can not readily adapt. In

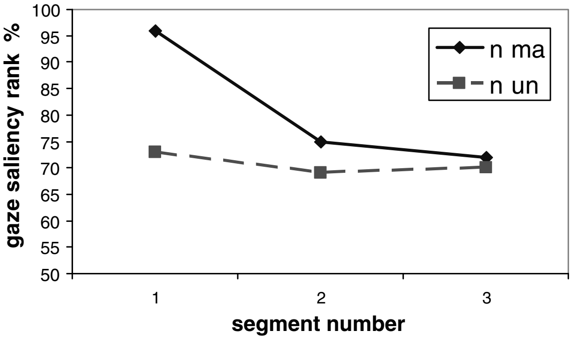

Figure 9. Change in model performance as function of distance from the training scene. This model, a raw patch model of high dimensionality ($L = 11$), is trained on subsets of scene A. We take A and divide it into three segments and train on segment 1. The highest matched performance occurs in segment 1, as expected. We note that as we move our testing segment away from segment 1, matched performance decreases with little change to unmatched performance.

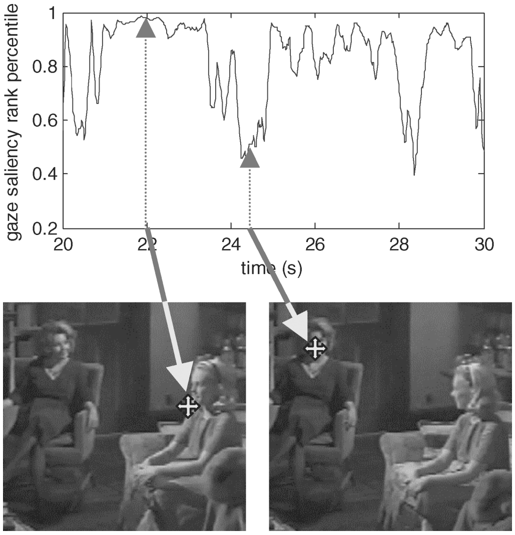

Figure 10. *Effects of Context on Models.* The top graph represents the time varying gaze saliency rank percentile computed for a human subject (under the 3 × 3 × 3 × 2 Itti Model) applied to his own trajectory. The arrows point to the actual visual scene shown at the associated point in time. The crosses represent the locations where the human subject was actually looking at those times. Note that the model is high-scoring at first, implying that it is well matched to the situation where the blonde rightmost female character is speaking. When the focus of attention of the human subject shifts to the left female character, the model is unable to account for this change.

some sense, knowing who is being talked to represents a complex social phenomenon: a truly high-level top-down effect. Our framework thus provides for mechanisms where the weaknesses in a particular visual attention model can be pinpointed and investigated.

Finally, we note that, within our framework, the more complicated Extended Itti Model does not necessarily perform any better, after tuning, than much simpler feature extraction methods. In some ways, this is not unexpected, since biologically-inspired models are not necessarily models that seek to replicate human gaze patterns, but rather are often intended to provide some didactic or theoretical role. Still, it is surprising how well a simple method, such as a set of Gaussian pyramid features, can perform even at low dimensionalities (Table 1, A on B, 4th row).

6. Discussion

Our system addresses the problem of how computational models of visual attention can be compared with human subjects, and thereby be compared with one another. Validation against human subjects is obviously not the only measure by which computational models of attention may be judged. Draper and Lionelle (2005), for example, evaluate the Itti Model in terms of its sensitivity to similarity transforms. Though Draper and Lionelle frame their investigation in terms of appearance-based recognition systems, their work is applicable more generally. The possibility that known statistical and theoretical properties of the human visual attention system be used to directly evaluate computational models is both intriguing and promising.

The use of random models as controls is one way that such properties could be investigated. The random models used in this current study all share one common aspect: they are computed without regard to absolute spatial and temporal information. Different choices of models which incorporate more information could help determine how particular aspects of the scene interact with the chosen features. For instance, we could randomly choose spatial locations from the set of gaze positions reported in human observers. Such a model would be spatially correlated but temporally uncoupled. Its use as a control would give an indication of the feature dependence on spatial versus temporal information. We could also use human subjects, perhaps engaged in specific tasks, such as target search, as a comparison against the free-viewing experiments we have seen here. Such search-based task patterns would be completely physiological, but the scanning patterns would represent a different underlying motivation. The teasing apart of specific interactions using appropriate controls is a current area of active investigation.

We should also note that though our formulation is based upon probabilistic intuitions, it does not serve necessarily as a generative model for visual attention. In other words, our computational framework is capable of revealing insights regarding how well a model is performing, but it makes no statement regarding what gaze policy should be applied.

An issue that makes it difficult to step directly to some generative model for gaze trajectories in our framework is the fact that visual attention is not stateless. Viewing visual attention as a purely feature-based probabilistic problem leads to behavior that is non-physiological. As seen in Fig. 11, human eye movements are composed of fixations interspersed with saccades. If we sample from an approximation to the underlying probability distribution, we ignore the strong temporal and spatial correlations inherent to human eye trajectories. It is likely that this framework could

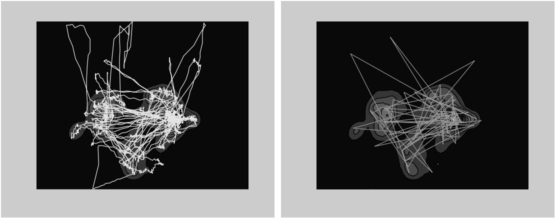

Figure 11. Human gaze data (left) and a trajectory drawn probabilistically from an approximation to the underlying density (right). Note that whereas saccades and fixations are identifiable in the left image, these properties do not exist in the right image.

benefit from some type of state, as would be found in a Markov model, for example.

As we have seen, another problem that complicates our analysis is the presence of context-dependent behavior. It is likely that an observer viewing some scene is constantly changing his preferences and objectives, as dictated not only by the scene, but also by some unobservable internal mental state. These shifting priorities and desires are likely one factor that contributes to the degradation of our saliency computation as the tested scene becomes temporally further removed. An alternate interpretation of the same effect is one of overfitting. However, if this were the case, the true unmatched normal to normal comparison would be better than what we have reported. As it stands, the results are already shown to be significantly different from scene-uncorrelated random models. We should note, however, that the issue of context and top-down effects somewhat hinges on definition. Context effects, in a computational sense, are those effects not adequately represented by the features of a given model. As these features come to be incorporated into a model, their validity as well as the extent of their applicability increases; correspondingly, the rapidity of performance loss due to shifts away from the training source decreases. A model that seeks to represent the scanning pattern of an observer examining a pair of faces laid side by side might have abysmal performance until it incorporates the fact that one of the faces is the mother of the observer.

Our lack of attention to local trajectory statistics and internal mental state is also reflected in our decision to omit the inhibition-of-return mechanism from the Itti Model, possible making its comparison an unfair one. However, it is not clear how inhibition-of-return could be adapted to a dynamic environment with motion, however, since the addition of a temporal component might suggest that the areas corresponding to inhibited behavior should be time-varying. In addition, a question arises as to how the inhibition-of-return mechanism should be initialized, as the gaze trajectory predicted by the model would interfere with future saliency calculations. This added complexity is likely to be partially why inhibition of return is omitted from many recent investigations of computational saliency (Carmi and Itti, 2006; Itti, 2005b; Itti, 2006b). However, we must admit that use of inhibition of return in the Itti model, which provides some local state and memory, could impact our results, though it is not clear whether it would make the Itti model perform better or worse, or whether other feature extraction methods would benefit from a similar mechanism.

Our custom implementation likely differs in some ways from the implementation available at (ilab.usc.edu). Because the specifics of the Itti model are clearly defined, with the possible exception of motion, we found it more expedient to implement the model directly. This resulted in a large improvement in the ability of the Itti model to adequately describe the gaze patterns of human observers. The Itti model has evolved substantially from its inception, and it is likely that recent incorporations of signal rectification and local enhancement which, visually, give a more interpretable picture of the salience associated with a given modality, also lead to some loss of information that is not recoverable, and thereby not available for optimization at our classification stage. We have examined the Itti model from multiple angles, under multiple testing conditions, and our results are similar in all permutations.

We should note that we have chosen one particular path in our framework for reasons of computational expediency and illustrative use, but many options exist. In particular, we have used the notion of saliency as an intermediary step in calculation mainly due to its intuitive nature. However, we are, in fact, evaluating trajectory generators simply by dimensionality reduction over human trajectories—a notion that does not actually require either a true probabilistic underpinning or an explicit formulation of saliency in the manner of Koch and Ullman (1985). There exists an equivalence class of possible saliency schema, the nature, limitations, and capabilities of which we hope to investigate in the future.

7. Conclusions

We have presented a general technique for evaluating whether a computational model for visual attention behaves in a human-like manner by direct comparison with human subjects. We have shown that distance metrics in image space are insufficient for a general concept of proximity for visual attention, and have developed a classification strategy employing dimensionality-reduction that instead operates in feature space. This classification strategy not only provides a more standardized basis for a notion of salience, but also provides a common interface upon which different models of visual attention may be compared. We have taken a probabilistic version of this classification strategy and transformed it into a dimensionality reduction problem, opening up a broad area of possible inquiry.

By employing our framework, we have shown that the popular, biologically-inspired bottom-up Itti Model, though it serves as a cornerstone for many robotic implementations for visual attention, does not necessarily provide any advantage in terms of emulating human behavior.

In conclusion, we have demonstrated how computational models of visual attention can be developed, applied, optimized, and evaluated: steps we hope bring us closer to robots that look at where humans look, on the road towards seeing what humans see.

Appendix A—The Extended Itti Model (Motion Extension)

Simple image difference detection (by computing the absolute difference between adjacent frames, as is done in Niebur and Koch (1996)) is insufficient as a basis for a motion modality, as it fails to highlight known pop-out effects (Fig. 12, right) (Wolfe, 2004). Similarly, employing optical flow (see (Beauchemin and Barron, 1995) for a review) as a basis for motion salience typically involves a retrospective interpretation of the optical flow field, a paradigm that does not fit neatly into the feedforward framework. Optical flow techniques which could be easily adapted, such as Heeger's work with hierarchical spatiotemporal Gabor filters (1988), are computational expensive as they incorporate numerical optimization at each image location.

We employ a compromise approach as a basis for computing motion saliency, a variation of time-varying edge detection (as recounted in Jain et al. (1995)). The time-varying "edginess" of a point, $E_{s,t}$, is computed as the product of the spatial and temporal derivatives:

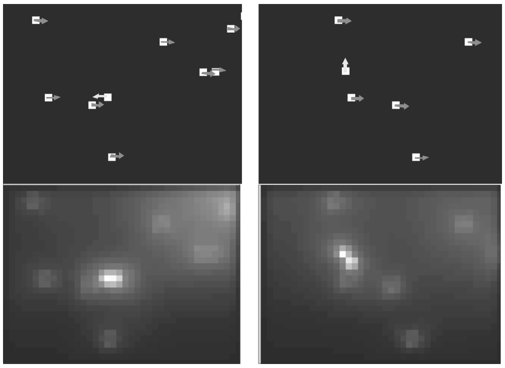

Figure 12. Motion pop-out stimuli composed of boxes (top panes) and associated final motion conspicuity map (bottom panes). The left figures represent directional competition, with a single stimuli moving leftwards in a field of rightward moving distractors. The right figures represent cross-orientation competition, with a single upwards moving stimuli popping-out among rightward moving distractors. The arrows in the top panes are for illustrative purposes only, and did not appear in the actual movie.

$$E_{s,t}(x, y, t) = D_s I(x, y, t) \cdot D_t I(x, y, t)$$

where I is the intensity of an image at the spatial coordinates, x and y, and at the temporal coordinate, t, and s is some spatial direction $s = s(x, y)$. In our work, we approximate the spatial derivative with the imaginary component of the Gabor-filtered image obtained during the basic Itti Model extraction, and obtain the temporal derivative from image differencing after temporal filtering. Note that this technique can only provide the combined magnitude of motion and intensity and not the magnitude of stimuli motion alone. This flaw, however, is mitigated by the multi-scale aspect of the Itti Model. Our motion extension is very much in the style of the Itti model as it is (i) integrated in a fashion similar to that of the orientation modality and does not break away from the original model's methodology or framework, (ii) computational quick and easy in implementation, and (iii) capable of describing a wide range of pop-out motion phenomena. An updated relational diagram for the Itti Model is shown in Fig. 13.

We begin by extending the original Itti Model (Itti et al., 1998) equations in time (e.g. the intensity modality $I(\sigma)$ becomes $I(t, \sigma)$, the red-green feature map $\mathcal{RG}(c, s)$ becomes $\mathcal{RG}(t, c, s)$, etc.) Working purely with image intensities, under the assumption that motion is largely color-blind, for N frames $I(t, \sigma), t \in$

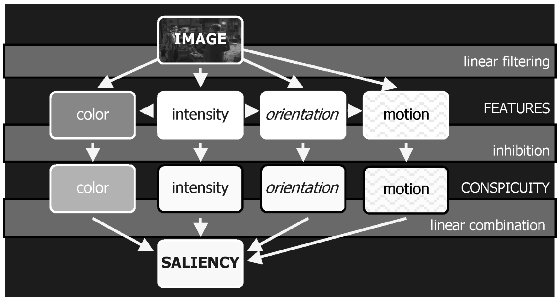

Figure 13. Relational diagram of Extended Itti Model. At the feature level, intensity informs orientation, which in turns informs motion. At the conspicuity level, aspects of each modality compete within that modality. The results of conspicuity computation are then funneled into a final saliency.

$[1 \ldots N]$, we obtain motion feature maps in the following manner:

(1) Compute the N-th order first temporal derivative, $M_t(t, \sigma)$.

(2) Compute the spatial derivative, $M_s(t, \sigma, \theta)$, from gradients extracted during orientation computation:

$$M_s(t, \sigma, \theta) = Im\{O_c(t, \sigma, \theta)\}$$

(3) Compute the motion feature map $\mathcal{M}(t, \sigma, \theta)$ as the product of $M_s(t, \sigma, \theta)$ and $M_t(t, \sigma, \theta)$:

$$\mathcal{M}(t, \sigma, \theta) = \mathcal{M}_s(t, \sigma, \theta) \cdot \mathcal{M}_t(t, \sigma, \theta)$$

The motion conspicuity map is derived directly from the above algorithm, using the normalization operator $\mathcal{N}$, and a cross-scale addition operator, $\oplus$, as defined in Itti et al. (1998), to emulate the effects of lateral inhibition:

(1) Compute the direction of motion for each orientation to obtain positive and negative directional features. The positive directional feature $\mathcal{M}_+(t, \sigma, \theta)$ is defined as $\sqrt{\mathcal{M}(t, \sigma, \theta)}$ at locations where $\mathcal{M}(t, \sigma, \theta)$ is positive, and 0 otherwise. Similarly, the negative directional feature $\mathcal{M}_-(t, \sigma, \theta)$ is defined as $\sqrt{-\mathcal{M}(t, \sigma, \theta)}$ at locations where $\mathcal{M}(t, \sigma, \theta)$ is negative, and 0 otherwise.

(2) Compute the directional contribution to motion conspicuity, $\mathcal{M}_d(t, \sigma, \theta)$ by allowing positive and negative directional motion features to compete locally:

$$\mathcal{M}_d(t, \sigma, \theta) = \mathcal{N}(\mathcal{M}_+(t, \sigma, \theta)) \oplus \mathcal{N}(\mathcal{M}_-(t, \sigma, \theta))$$

This accounts for popout phenomena such as shown in on the left in Fig. 12.

(3) Compute the across-scale contribution for each orientation, $\mathcal{M}_o(t, \theta)$.

$$\mathcal{M}_o(t, \theta) = \bigoplus_{\sigma=0}^{8} \mathcal{N}(\mathcal{M}_d(t, \sigma, \theta))$$

This is equivalent to saying that all scales at a particular orientation compete with one another.

(4) Compute the conspicuity map for motion, $\bar{M}(t)$, by combining across all orientations:

$$\bar{M}(t) = \sum_{\theta \in \{0, \frac{\pi}{4}, \pi, \frac{3\pi}{4}\}} \mathcal{N}(\mathcal{M}_o(t, \theta))$$

Motion conspicuity is then added, as an additional modality, to the final saliency map:

$$S = \frac{1}{4}\left(\mathcal{N}(\bar{I}) + \mathcal{N}(\bar{C}) + \mathcal{N}(\bar{O}) + \mathcal{N}(\bar{M})\right)$$

replacing the basic Itti Model equation for saliency (which only neglects a term for motion).

Acknowledgments

The authors would like to thank Warren Jones and Ami Klin for providing the human eye-tracking data used in this work. We are also indebted to Steven W. Zucker for his help and valuable advice. The authors are indebted to the comments of two anonymous reviewers who have provided valuable suggestions to this work. Support for this work was provided by a National Science Foundation CAREER award (#0238334) and award #0534610 (Quantitative Measures of social response in autism). Some parts of the architecture used in this work was constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots). This research was supported in part by a grant of computer software from QNX Software Systems Ltd.

References

Balkenius, C., Eriksson, A.P., and Astrom, K. 2004. Learning in visual attention. In *Proceedings of LAVS '04*. St Catharine's College, Cambridge, UK.

Beauchemin, S.S. and Barron, J.L. 1995. The computation of optical flow. *ACM Computing Surveys,* 27(3):433–466.

Breazeal, C. and Scassellati, B. 1999. A context-dependent attention system for a social robot. In *Proceedings of the Sixteenth international Joint Conference on Artificial Intelligence,* T. Dean (Ed.). Morgan Kaufmann Publishers, San Francisco, CA pp. 1146–1153.

Burgard, W., Cremers, A.B., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W., and Thrun, S. 1998. The interactive museum tour-guide robot. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial intelligence/innovative Applications of Artificial intelligence*, pp. 11–18.

Burt, P.J. and Adelson, E.H. 1983. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, COM-31(4):532–540.

Carmi, R. and Itti, L. 2006. Causal saliency effects during natural vision. In *Proc. ACM Eye Tracking Research and Applications*, pp. 1–9.

Draper, B.A. and Lionelle, A. 2005. Evaluation of selective attention under similarity transformations. *Computer Vision and Image Understanding.* 100:152–171.

Duda, R.O. and Hart, P.E. 1973. *Pattern Recognition and Scene Analysis*. John Willey: New York.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. 2003. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166.

Fujita, M. 2001. AIBO: Toward the era of digital creatures. *The International Journal of Robotics Research.* 20:781–794.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A., and Wang, J. 2005. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 2199–2204.

Gottlieb, J., Kusunoki M., and Goldberg, M.E. 1998. The representation of visual salience in monkey posterior parietal cortex. *Nature,* pp. 391:481–484.

Heeger, D.J. 1988. Optical flow using spatiotemporal filters. *IJCV.* 1:279–302.

iLab Neuromorphic Vision C++ Toolkit (iNVT). 2006. Retrieved June 5, 2006, from http://ilab.usc.edu/toolkit/home.shtml

Imai, M., Kanda, T., Ono, T., Ishiguro, H., and Mase, K. 2002. Robot mediated round table: Analysis of the effect of robot's gaze. In *Proc. 11th IEEE Int. Workshop Robot and Human Interactive Communication* (ROMAN 2002), pp. 411–416.

Itti, L., Koch, C., and 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.

Itti, L., Dhavale, N., and Pighin, F. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*.

Itti, L., Rees, G., and Tsotsos, J.K. (Eds.) 2005a. *Neurobiology of Attention.* Elsevier Academic Press.

Itti, L. 2005b. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123.

Itti, L. and Baldi, P. 2006a. Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, 19:1–8.

Itti, L. 2006b. Quantitative modeling of perceptual salience at human eye position. *Visual Cognition*, (in press)

Jain, R., Kasturi, R., and Schunck, B.G. 1995. *Machine Vision*. McGraw-Hill Science/Engineering/Math.

Klin, A., Jones, W., Schultz, R., Volkmar, F., and Cohen, D. 2002. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch Gen Psychiatry,* 59:809–816.

Koch, C. 1984. A theoretical analysis of the electrical properties of an X-cell in the cat s LGN: Does the spine-triad circuit subserve selective visual attention? *Artif. Intell. Memo* 787, MIT, Artificial Intelligence Laboratory.

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

Kustov, A.A. and Robinson, D.L. 1996 Shared neural control of attentional shifts and eye movements. *Nature*, 384:74–77.

Lee, D.K., Itti, L., Koch, C., and Braun, J. 1999. Attention activates winner-take-all competition among visual filters. *Nat. Neurosci.*, 2(4):375–381.

Li, Z. 2002. A saliency map in primary visual cortex. *Trends Cogn. Sci*, 6:9–16.

Mazer, J.A. and Gallant J.L. 2003. Goal-related activity in V4 during free viewing visual search: Evidence for a ventral stream visual salience map. *Neuron*, 40:1241–1250.

Nagai, Y., Asada, M., and Hosoda, K. 2002. Developmental learning model for joint attention. In *Proceedings of the 15th International Conference on Intelligent Robots and Systems (IROS 2002)* (Lausanne, Switzerland), pp. 932–937.

Niebur, E., Itti, L., and Koch, C. 1995. Modeling the "where" visual pathway. In *Proceedings of 2nd Joint Symposium on Neural Computation, Caltech-UCSD*, Sejnowski, T.J. (Ed.), Institute for Neural Computation, La Jolla, vol. 5, pp. 26–35.

Niebur, E. and Koch, C. 1996. Control of selective visual attention: Modeling the "where" pathway. *Advances in Neural Information Processing Systems,* In Touretzky, D.S., Mozer, M.C., and Hasselmo, M.E. (Eds.), MIT Press: Cambridge, MA, vol. 8, pp. 802–808.

Ouerhani, N., von Wartburg, R., Hugli, H., and Muri, R. 2004. Empirical validation of the saliency-based model of visual attention. *Elec. Letters on Computer Vision and Image Analysis*, 3:13–24.

Parkhurst, D., Law, K., and Niebur, E. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123.

Petersen, S.E., Robinson, D.L., and Morris, J.D. 1987. Contributions of the pulvinar to visual spatial attention. *Neuropsychologia*, 25:97–105.

Robinson, D.L. and Petersen, S.E. 1992. The pulvinar and visual salience. *Trends Neurosci.*, 15(4):127–132.

Salvucci, D.D. and Goldberg J.H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on Eye tracking research & applications*, pp. 71–78.

Scassellati, B. 1999. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Lecture Notes in Computer Science*, 1562:176.

Tatler, B.W., Baddeley, R.J., and Gilchrist, I.D. 2005. Visual correlates of fixation selection: Effects of scale and time. *Vision Res*, 45(5): 643–659.

Tessier-Lavigne, M. 1991. Phototransduction and information processing in the retina. In *Principles of Neural Science*, E. Kandel, J. Schwartz, and T. Jessel (Eds.), Elsevier Science Publishers B.V. pp. 401–439.

Torralba, A. 2003. Modeling global scene factors in attention. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 20:1407–1418.

Treue, S. 2003. Visual attention: The where, what, how and why of saliency. *Curr Opin Neurobiol*, 13(4):428–432.

Tsotsos, J.K. 1988. A 'complexity level' analysis of immediate vision. *International Journal of Computer Vision (Historical Archive)*, 1(4):303–320.

Tsotsos, J.K., Culhane, S.M., Winky, Y.K.W., Yuzhong, L., Davis, N. and Nuflo, F. 1995. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1):507–545(39).

Tsotsos, J.K., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., and Zhou, K. 2005. *Attending to Motion, Computer Vision and Image Understanding*, 100(1–2):3–40.

Turano, K.A., Geruschat, D.R., and Baker, F.H. 2003. Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43(3):333–346.

Wolfe, J.M. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review.*, 1(2):202–238.

Wolfe, J.M. and Gancarz, G. 1996. Guided Search 3.0: A model of visual search catches up with Jay Enoch 40 years later. *Basic and Clinical Applications of Vision Science*. Kluwer Academic: In V. Lakshminarayanan (Ed.), Dordrecht, Netherlands.

Yee, C. and Walther, D. 2002. Motion detection for bottom-up visual attention, tech. rep., SURF/CNS, California Institute of Technology.