ON THE ROLE OF CONTEXT IN PROBABILISTIC MODELS OF VISUAL SALIENCY

Neil D. B. Bruce, Pierre Kornprobst

INRIA, 2004 Route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex, France

ABSTRACT

In recent years, many principled probabilistic definitions for the determination of visual saliency have been proposed. Moreover, there has been increased focus on the role of context in the determination of visual salience. Prior efforts have shed some light on how context may help in predicting the location of, or presence of features associated with an object in the context of detection or recognition. Nevertheless, there remains a variety of manners in which context may be exploited towards providing better judgements of salient content. In this light, we investigate the role of context in the probabilistic determination of salience while presenting a number of potential avenues for future research.

Index Terms— saliency, context, image statistics, attention

The complexity of visual search demands strategies for focusing high-level processing on some subset of the incoming stream of visual input at the expense of detailed processing of other visual input [1]. This focal processing may take the form of biased processing towards certain locations or features in the scene. At the same time, outside of the demands of a particular visual task definition, it is important to be alerted to content that may be of interest in its own right, for example a predator suddenly appearing while an animal is searching for food. These two elements constitute respectively, the task driven top-down side of attention which serves to instigate bias towards task relevant content, and the bottom-up side which may be viewed as a stimulus driven component which results in the deployment of attention towards conspicuous visual patterns. Recently, a variety of models of saliency and attentional bias have emerged having as a basis a probabilistic definition for content of interest. There are a number of elements in the computation performed by these models that differ from one model to another and that are important as they impact on the behavior of the models. Moreover, there are a variety of issues that relate to the notion of *context* in probabilistic saliency computation that deserve further consideration. This is in essence the subject matter of this paper; while the subject matter put forth demonstrates the efficacy or importance of context in certain aspects of saliency computation, this is also equally a *road map* indicating a variety of promising avenues for further research efforts and in addition, strategies that may be exploited depending on the nature of the task under consideration.

The structure of the paper is as follows: We first begin with an overview of recent models of visual saliency computation that have at their core, a probabilistic determination of saliency. This includes some discussion of the differences between these proposals and additionally highlights areas where contextual information has been successfully exploited to improve the explanatory power of the models in question. Following this, we consider a few important issues pertaining to the determination of visual salience. These are respectively, the role of location in saliency computation, and the role of environmental statistics in the determination of saliency.

1. BACKGROUND

In recent years, a variety of proposals for the computation of visual saliency have emerged which form judgments of saliency on the basis of a probabilistic determination. In this section, we provide an overview of these proposals and highlight areas in which contextual information is currently employed for the purposes of saliency computation.

1.1. An Information Theoretic Approach

In [2, 3], the authors propose a strategy for visual saliency computation based on an information theoretic approach deemed attention based on information maximization (AIM). The authors propose a strategy for the determination of visual saliency that is analogous to Shannon's work on the transmission of English words [4]. In short, the salience of a local neighborhood x of the scene is given by its self-information $-log(p(x|C))$ where C is the context on which this estimate is based. In [2] it is suggested that this context may be a local neighborhood surrounding the local observation x, but is computed with C constituting the entire scene for computational parsimony. The likelihood estimate in this case is achieved through the use of a set of filters learned through Independent Component Analysis (ICA). This results in a set of feature maps that may be assumed statistically independent and follows the proposal made in [5]. This operation reduces the likelihood estimate from one in a $3N^2$ dimensional space (with N the width of a local patch in RGB space), to $3N^2$ one dimensional density estimates. This is an important contribution as it places the likelihood estimate of a local patch within a form that is computationally tractable.

1.2. A Discriminant Approach

In [6], saliency is formulated within the context of a discriminant definition. This amounts to considering the power of some set of features to discriminate between observations drawn from a central region and those drawn from a surrounding region. Specifically, given some set of features $\mathbf{X} = X_1, ..., X_d$, a location l and a class label Y with $Y_l = 0$ corresponding to samples drawn from the surround region and $Y_l = 1$ corresponding to samples drawn from a smaller central region centered at l. The judgement of saliency then corresponds to a measure of mutual information, computed as

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement no. 215866, project SEARISE.

$I(\mathbf{X}, Y) = \sum_{i=1}^{d} I(X_i, Y)$. Note that there are once again some computational tricks employed to make the overall estimate $I(\mathbf{X}, Y)$ computationally tractable. In this case, total independence is not required, but it suffices to assume that considering pairwise combinations of features does not help appreciably in the discrimination problem. It is also worth noting that the local estimate of the distributions of features in $\mathbf{X}$ conditioned on Y are assumed to fit a Generalized Gaussian distribution for computational parsimony.

1.3. The Bayesian Strategy

In [7, 8], saliency is formulated on the basis of Bayes rule. The definition differs slightly in these two efforts in that in [7] a location prior for a particular object is formed on the basis of the response of global receptive fields in line with a *gist* view of processing. In [8] judgements are based solely on local responses conditioned on the statistics of natural images, a proposal first appearing in [5]. The following formulation is based on the latter of these studies for the purposes of exposition, and is similar in nature to that put forth in [7]. In short, the quantity considered is: $s_z = p(C = 1|F = f_z, L = l_z)$ where s_z indicates the saliency of coordinate location z, $C = 1$ implies membership to a particular class, $F = f_z$ reflects the responses of receptive fields observed at location z and $L = l_z$ corresponds to the location in question. It is then further suggested that the quantity $p(F = f_z, L = l_z)$ can be computed by considering the product of the marginals $p(F = f_z)p(L = l_z)$. For convenience, locations are compared on the basis of their log likelihood, yielding a final expression of: $-log(p(F = f_z)) + log(p(F = f_z)|C = 1) + log(p(C = 1|L = l_z))$. The first of these terms it is noted corresponds to the bottom-up measure of saliency akin to that put forth in [2]. The latter terms provide a measure of the likelihood of observed features conditioned on the class of interest and a prior on location.

1.4. Surprise

In the Surprise model [9], saliency is computed as the distance between a prior model M (e.g. a distribution of responses based on a particular set of features) and the posterior probability distribution given new data $P(M|D)$. The Kullback-Leibler divergence yields a measure of the extent to which the observation based on the new data is surprising taking into account the prior.

2. EXPLOITING CONTEXT: EXISTING STRATEGIES

While the literature currently presents a handful of strategies aimed at exploiting contextual information, there remains a variety of unexplored avenues for further contribution of context towards the determination of salience. As the proposals we have discussed differ in the definition of the context on which the computation of salience is based, this is evidently an important issue. In this paper, we discuss a variety of issues pertaining to the determination of saliency specifically insofar as context impacts on the likelihoods involved in its determination. As such, rather than being a proposal for a specific strategy for saliency computation, we instead focus on more general issues that are important in light of the models discussed. These issues will be important for any model of saliency based on a probabilistic formulation. This discussion presents a wide array of possibilities for future research efforts and also serves to highlight some subtle but important issues in probabilistic saliency computation.

2.1. Priors based on Location

The use of location information appears prominently in the two approaches that are based on a Bayesian formulation. In [7] the location conditioned on the global receptive fields provides a prior on the location of an object of interest. This is shown to be a powerful strategy with the contextual determination of object location providing a stronger predictor of object location than a measure of bottom up saliency. In this paper, we are concerned with how contextual knowledge in a general sense may factor into the bottom-up determination of visual saliency including both positional and feature based knowledge. To give a more concrete example, in the model put forth in [8], the term $p(F = f_z, L = l_z)$ appears, which the authors suggest may be considered as two independent terms $p(F = f_z)p(L = l_z)$. There are two issues pertaining to this term that are worthy of discussion. The first issue concerns the use of location as an independent factor in determining target saliency.

It has been demonstrated in two previous studies [8, 10] that fixation data contains a strong central bias. On the basis of this, it is suggested in [8] that the location prior in itself may be useful in the overall determination of saliency. That said, in a recent study [11], it has been shown that this observation appears to be merely an artifact of the fact that images are composed, or that they are presented on a computer monitor, and does not appear in general in free visual sampling.

Therefore the following might be said: If the task at hand is one of detecting items of interest in an image in which the photographer has centered a target of interest, which may well be the case in an image processing context, then it is suitable to use a location based prior in determining saliency as this may well improve the judgment of salient content appreciably. In fact, this cue in itself may be more useful than the state of the art in saliency judgments as evidenced in [8] and [10]. If the task in question is to construct an accurate model of human judgements of saliency or consists of a module in say, a mobile robot navigation system, it is inappropriate to employ such a strategy as this cue does not inform on the presence of salient content in a general sense.

A second issue that is also of interest from a modeling perspective, is the consideration of likelihood in which features and location are considered jointly. That is, there is in some instances clearly an *a priori* expectation of certain features tied to location. Consider for example a football match in which the players and action are largely confined to one end of the field. An overzealous fan running on to the pitch at the opposite end of the field may well result in the direction of gaze to this end of the field even in the event that the running behavior of the fan is not significantly different than that of the players at the opposing end of the field. From a more applied perspective, in a surveillance system one is likely to have strong priors concerning what sort of features are expected within different areas of the visual field. For example, movement around a door to a restricted area, or around a fenced area should be judged more salient than movement on a common pathway to the refreshment stand. It is therefore sensible to consider gains that may be had in considering features and location jointly.

In this work, we have considered the extent to which a combined feature/location prior might predict suspicious activity in the

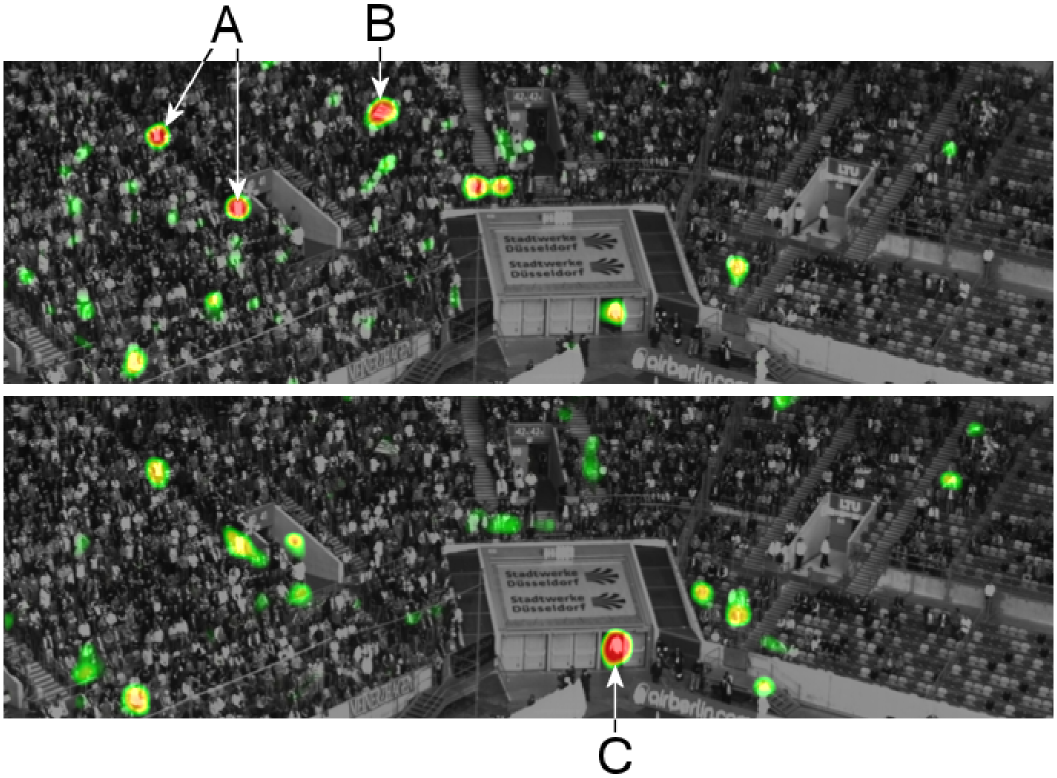

Fig. 1. Saliency determination based on the SEARISE stadium video surveillance sequence: Top: Saliency given by $-log(p(x))$ with response likelihood based on the current response of filters oriented in space-time computed over the entire scene. Bottom: Response likelihoods are associated with each pixel location with likelihoods based on a temporal support. Note the significant difference in computed salience: In the global estimate, a quickly waving flag (B) is the most salient target. In the temporal case, a high degree of motion has been observed at this location over time and instead a more subtle movement of a man entering a doorway where movement is unexpected is deemed most salient (C). The relative saliency of people moving along a more common pathway also differs across conditions (A).

context of fixed video from a football stadium (See Figure 1). Features employed are based on independent components learned using the Infomax ICA algorithm [12], with PCA preprocessing preserving 95% variance (60 components). For each location in the stadium video, the joint likelihood $p(F = f_z, L = l_z)$ is observed explicitly $\forall z$. Subsequently, the judgment of saliency is given by $-log(p(F = f_z, L = l_z))$ in contrast to previous efforts that consider the marginal likelihoods only [8]. An example of this operation is shown in figure 1 (bottom). The relatively subtle movement of a man appearing in a lower doorway to the stadium (C) is assigned a high level of saliency on the basis that any activity within this region is unexpected. In contrast, with salience computed based on global motion patterns currently being observed (top) the highest degree of salience is assigned to a fast moving flag (B). One can imagine this sort of analysis to be useful in the detection of a range of important events such as illegal turns by vehicles, detection of items left behind, detection of unexpected events and analysis of crowd behavior as exemplified here. In practice, saliency computation that leverages information derived from several levels of contextual abstraction may be especially promising.

2.2. Exploiting Properties of the Environment

The proposal for the scale at which saliency computation takes place varies within the studies we have described, from a local surround region [2, 3, 6] to the entire image [2, 7], to the space of all natural images [5, 8]. This is a consideration that no doubt impacts upon the resultant judgements of saliency and is deserving of further consideration. An additional level at which the determination of saliency may be computed is at a level somewhere between that of the current scene and the space of all natural images. For example, if one is walking in a forest, one has certain expectations concerning the content of the surrounding region and this impacts on prior expectation. This is exemplified by Figure 2 in considering the image of a car lying in a forested region. Figure 2 depicts a montage of urban and forested images that were used to learn representative statistics of each environment. Distributions were learned based on a 100 bin histogram density estimate for a set of local filters based on ICA [12] as in the example shown in Figure 1. Below this are some example images and a depiction of how environmental statistics may lead to different judgements of salience. The details are as follows: We have conducted some simple experiments to determine the extent to which *environmental statistics* might be employed to provide a stronger judgement of salient content. If one considers

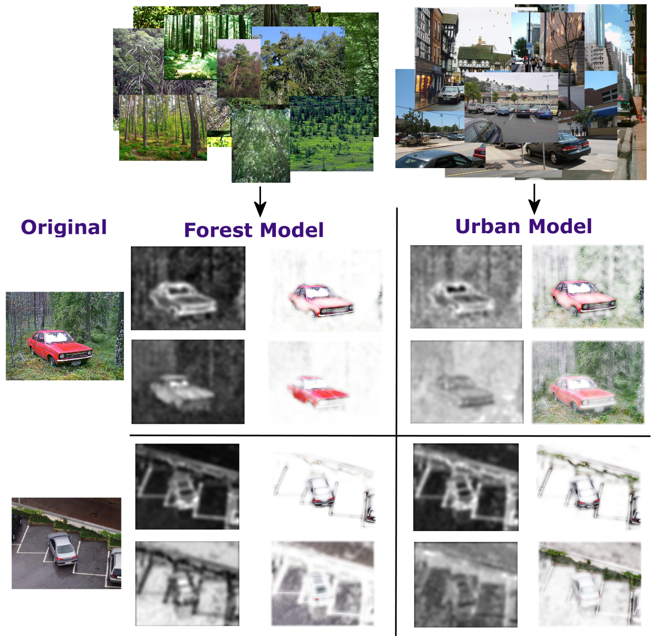

Fig. 2. Top: Training samples used to learn a representation of forest and urban statistics respectively. The four quadrants show examples of output corresponding to the two images appearing on the left and corresponding to the two categories of statistics appearing above. Within each quadrant: Top left: $-log(p(F = f_z|context))$. Top right: Original image modulated by $-log(p(F = f_z|context))$. Bottom left: $-log(p(F = f_z|context)/p(F = f_z|N))$. Bottom right: Original image modulated by $-log(p(F = f_z|context)/p(F = f_z|N))$.

simply the quantity $p(F = f_z|forest)$ versus $p(F = f_z|urban)$, there exists some subtle differences in the resultant saliency maps. This is depicted in the top row of each quadrant of Figure 2. Note that training based on a forest relative to an urban environment yields relatively more confidence for the foliage based on the urban statistics. In the second image example, the overall scene is more consistent with an urban environment and the same trend is observed. Differences may be seen more clearly in considering a quantity that reflects the belief that one is within a particular environment based on what is observed. For example, if one considers the quantity $p(F = f_z|forest)/p(F = f_z|N)$ (where N denotes the space of all natural images), significant differences in output emerge. This quantity might additionally be viewed as the reciprocal of the extent to which observed responses predict the environment one finds themselves in. Figure 2 demonstrates the output of this quantity on a log scale and in addition, the original image modulated by the measured salience (with less salient regions appearing more white) on the bottom row of each quadrant.

It is worth noting that the notion of prior expectation need not be categorical, but might also take on a more relaxed definition such as a prior based on recent experience, or task definition. Additionally, a rich contextual model might include an assessment of likelihoods based on various different levels of contextual abstraction with all levels contributing to the assessment of perceptual salience.

3. DISCUSSION

We have highlighted a variety of important issues pertaining to the role of context in the probabilistic determination of saliency. First, the consideration of likelihoods of features tied to location in a scene may be a strategy that holds promise for various computer vision or image processing applications. We have also demonstrated how contextual information might be leveraged to produce stronger priors on expected observations. All of the considerations discussed are amenable to inclusion within any of the probabilistic proposals for saliency computation that are reviewed. Additionally, there are many observations made that have not been tested explicitly, but nevertheless provide some important possible avenues for future research efforts in the probabilistic determination of visual saliency.

4. REFERENCES

[1] J.K. Tsotsos, "A complexity level analysis of immediate vision," *International Journal of Computer Vision*, vol. 2, pp. 303–320, 1988.

[2] N. D. B. Bruce and J.K. Tsotsos, "Saliency based on information maximization," *Advances in Neural Information Processing Systems*, vol. 18, pp. 155–162, 2006.

[3] N. D. B. Bruce and J.K. Tsotsos, "Saliency, attention and visual search: An information theoretic approach," *Journal of Vision*, vol. 9(3):5, pp. 1–24, 2009.

[4] C.E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[5] N. D. B. Bruce, "Image analysis through local information measures," *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 1, pp. 616–619, 2004.

[6] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," *Advances in Neural Information Processing Systems*, vol. 20, 2008.

[7] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search," *Psychological Review*, vol. 113(4), pp. 766–786, 2006.

[8] 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(7):32, pp. 1–20, 2008.

[9] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Advances in Neural Information Processing Systems*, vol. 19, pp. 1–8, 2006.

[10] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[11] F. Schumann, W. Einhuser-Treyer, J. Vockeroth, K. Bartl, E. Schneider, and P. Knig, "Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments," *Journal of Vision*, vol. 8(14):12, pp. 1–17, 2008.

[12] T-W. Lee, M. Girolami, and T.J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources," *Neural Computation*, vol. 11(2), pp. 417–441, 1999.