2012 Ninth Conference on Computer and Robot Vision

Recurrent Refinement for Visual Saliency Estimation in Surveillance Scenarios

Neil D. B. Bruce*, Xun Shi*, and John K. Tsotsos
Department of Computer Science and Engineering and
Centre for Vision Research
York University, Toronto, ON, Canada
4700 Keele Street, M3J 1P3
{*neil,shixun,tsotsos*}*@cse.yorku.ca*

***Abstract*—In recent years, many different proposals for visual saliency computation have been put forth, that generally frame the determination of visual saliency as a measure of local feature contrast. There is however, a paucity of approaches that take into account more global holistic elements of the scene. In this paper, we propose a novel mechanism that augments the visual representation used to compute saliency. Inspired by research into biological vision, this strategy is one based on the role of recurrent computation in a visual processing hierarchy. Unlike existing approaches, the proposed model provides a manner of refining local saliency based computation based on the more global composition of a scene that is independent of semantic labeling or viewpoint. The results presented demonstrate that a fast recurrent mechanism significantly augments the determination of salient regions of interest as compared with a purely feedforward visual saliency architecture. This demonstration is applied to the problem of detecting targets of interest in various surveillance scenarios.**

***Keywords*-attention; saliency; targeting; recurrence; information theory; surveillance; computer vision; visual neuroscience**

I. Introduction

Attention is arguably an indispensable component in attempting to solve the general problem of vision. One of the central arguments for attention is that it provides a solution to overcoming the computational complexity of visual search [12]. Models from the vision literature [10] also appeal to the importance of attention in the integration of separated features into a single unified representation. There is a multitude of ways in which attention may manifest itself in the context of a computational vision system.

Computational models of visual saliency [5], [2], [8] predict which items or locations in a scene are likely to draw an observers attention or gaze. The predictions of these models are generally independent of task directives, or of semantic contextual information. This is an instance where the role of attention appears in a bottom-up, feature driven fashion.

In contrast, there exist models that are concerned with the routing of information through a hierarchical representation of visual information [11], [7]. In some instances one can draw direct parallels between the routing of information and the problem of computational complexity in vision [12].

In [9], the authors propose a model of scene *gist*, that provides predictions of visual regions of interest and that augments the determination of salient regions with a more holistic global representation of scene content. A central element of this proposal, is the role that a global representation of the scene plays in predicting regions of interest. This proposal posits that global receptive fields provide a coarse representation of the scene that augments the bottom-up determination of salient regions. This requires training based on scene labels and a prior model for each class.

In this paper, inspired by computational mechanisms that appear in the primate brain, we propose a model that includes yet another specific but important mechanism: The role of fast recurrent loops in facilitating visual computation. Following in the tradition of *gist* based attention modeling, we examine the behavior of this mechanism insofar as its ability to augment the determination of visual saliency is concerned. It is worth noting that the conclusions and computational mechanisms discussed may be generalized to virtually any visual task that includes a spatiotemporal scale-space decomposition of the scene. The common underlying theme across these models lies in using large scale scene information to direct more specific local processing. The important contribution of this work is the novel recurrent processing that is used to augment the underlying representation in a manner different from that of direct feedforward computation or based on the global scene envelope or label as in the *gist* model. Unlike the gist based model, the proposed mechanism does not require any *a priori* training, or human labeling of data.

As a whole, the paper demonstrates the important role that recurrence may play from the perspective of attention and task directed processing. With that in mind, the paper is structured as follows: In section II, we discuss in more detail existing accounts of global-local processing in modeling attention, including the *gist* based account of global scene representation. Following this, we discuss in detail an important theory of the role of recurrence in processing based on the primate brain that motivates the modeling work put forth in this paper. This theory, termed the *fast brain*

* N. D. B. Bruce and X. Shi contributed equally to the content of this manuscript.

978-0-7695-4683-4/12 $26.00 © 2012 IEEE
DOI 10.1109/CRV.2012.23

theory derives from careful measurement of latencies tied to different visual pathways and points to an important role of recurrence in the corresponding visual representation. In section III, the computational details of our model including fast recurrence and saliency computation are outlined. In section IV, the model is validated in comparing the output that results with and without recurrent modulation according to a definition of visual saliency that has been shown to perform well in highlighting targets of interest or predicting human eye movements [13]. Finally, in section V, we discuss implications of the work appearing in this paper, as well as highlighting some promising avenues for further consideration.

II. On the *global* representation of a scene

In the machine vision literature, there are few efforts that recognize the role of a global or holistic representation of a scene in guiding visual tasks such as attention. It is apparent from existing computational efforts in this domain, that significant gains may be had in exploiting more global scene information for attention or other visual tasks [9]. In biological vision, the forest before trees precedence is well established based on a large body of psychophysical experiments [6]. In this section, we discuss in more detail the gist based account of scene representation and its role in attention. Following this, we present an argument from the biological vision literature for the structure of processing in visual pathways that emphasizes timing, and in particular, the crucial role that recurrence plays in a visual processing hierarchy. The computational modeling work presented in this paper derives from this *fast brain* theory and the experimentation establishes the importance of this computation within a modeling context.

A. Scene Gist

The central claim of the gist account of scene representation is that a global statistical representation of an image may aid in the prediction or localization of regions of interest. The support for this claim comes in the form of a body of recent psychophysical experiments that reveal that certain inferences may be made concerning a visual scene based on viewing an image presented for only a very short time course [4]. The model is such that a scene is analyzed on the basis of a set of *global* receptive fields. These global receptive fields are constructed based on the pooling of spatially separated Gabor filters, spaced evenly on a grid and subjected to a dimensionality reduction step via PCA. The bottom-up representation of saliency is augmented by a measure of the likelihood of the object appearing at each location (considering only the vertical coordinate), given the global representation.

The global representation presented by this model may modulate the overall determination of saliency as a function of the likelihood of spatial position in the scene conditioned on the global representation and tied to a particular object class. There are some inherent limitations to such a strategy in that it assumes a particular viewpoint, and scene composition. In addition, there are conceivably cues that derive from the global composition of a scene that are not tied to either position or the identity of a particular object. This begs the question of whether there may exist some more general mechanisms that allow more global scene analysis to bias further specific localized processing.

B. Fast brain theory

Visual computation in the brain proceeds along two separate pathways, deemed the ventral and dorsal pathways. This division first appears at the level of the retina as type M ganglion cells respond to lower contrast stimuli and quickly propagate action potentials. P type ganglion cells are characterized by slower activation and respond to local high contrast intensity variations. M cells project to the dorsal pathway which includes several visual areas characterized by large receptive fields that respond more strongly to low spatial frequency achromatic structure and high temporal frequencies and includes several important areas that represent motion. P cells project to the the ventral pathway, which represents form and color information and is characterized by smaller receptive fields that are chroma sensitive and that respond to precisely localized high spatial frequency patterns.

The nature of the interaction between these pathways and the timing of information flow has been elucidated by Bullier who measured the latencies of neurons among different visual areas [3] following visual stimulation. Different conduction velocities tied to neurons along the dorsal and ventral pathways implies that information flows at different rates dependent on the pathway under consideration with the dorsal pathway exhibiting faster conduction velocities than the ventral stream. Importantly Bullier showed that high level dorsal areas become active before low level ventral areas and that early ventral areas are modulated based on activity among higher dorsal areas approximately 20 ms before feedforward input reaches these early ventral areas [3]. This implies that a mechanism of fast recurrent refinement of ventral information by dorsal information is in action. It is important then to consider the nature and role of this interaction and also to consider what gains may be had regarding system behavior in including such a mechanism in a computational model.

In line with what has been discussed in the context of shortcomings of the gist approach, this provides an implicit means for more global scene structure to impact more local elements of a scene in rapid fashion, and in a manner not tied to vertical position or object or scene category. The impact of this interaction may include a wide range of strategies for recurrent modulation based on the specific connectivity

between and properties of the computational units that make up the two pathway visual hierarchy.

III. FRAIM: Fast Recurrent Attention by Information Maximization

The proposed model is produced based on the addition of hierarchical computation and recurrent connectivity to an existing model of bottom-up visual saliency computation [2]. The use of this model is based on two separate considerations: The extension of this model is a natural choice owing to the fact that saliency computation is determined by the underlying activation of filters. Altering the representation carried by these filters has the side effect of changing the resulting determination of salience. This is important as it allows the same measure of salience to be computed in either the presence or absence of recurrent modulation on the putative visual representation on which the determination of salience is based. Unlike the gist based approach where the role of global scene representation produces an independent decision regarding salience that is combined via multiplication with the bottom-up determination, the role of recurrence on the ventral representation has an implicit effect on the resulting determination of salience in that recurrence merely modulates the response of units that form the ventral representation. Subsequently saliency is determined based on the refined representation producing output that is implicitly biased by recurrent modulation. The second choice for this particular definition of visual saliency, is its simplicity and apparently strong performance in recent quantitative comparisons [13] with other models. In this manner, we can be certain that the augmented algorithm performance produces a model that outperforms the existing body of visual saliency algorithms. With this in mind, the remainder of this section describes the details of the implemented model.

The model consists of a hierarchical configuration of visual areas inspired by the human visual system.

A depiction of the overall model appears in figure 1.

The first visual layer is marked LGN. As may be seen in figure 1, there is an immediate functional division of the visual input into two separate LGN processing streams. These correspond to the dorsal and ventral pathways previously mentioned in discussing the fast brain theory. The distinction at this stage in the division of LGN into dorsal and ventral units is determined by the spatial and temporal frequency coding properties of the constituent filters. This is such that the dorsal stream consists of units on the low spatial and high temporal frequency end of the spectrum, and ventral units corresponding to high spatial and low temporal frequencies.

The LGN cells have a circular-symmetric receptive field profile and are modeled as 2D Difference-of-Gaussian filters. The structure of LGN filters is given by:

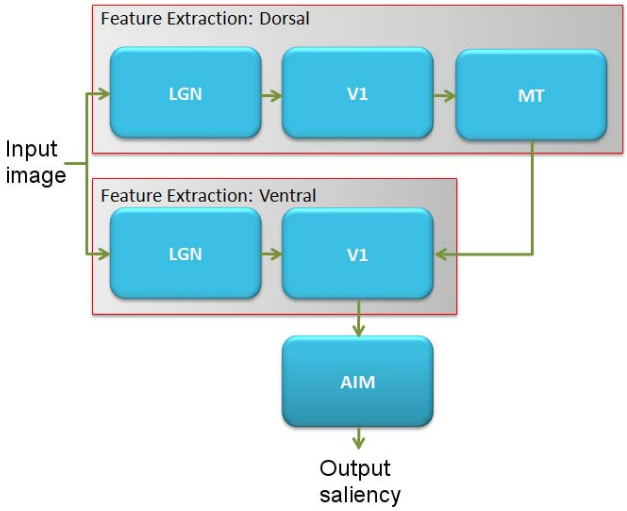

Figure 1. A schematic of the model demonstrating the various visual representations and connectivity appearing in the model. The visual input is first represented by LGN which responds to spatial and temporal variations in signal in a non orientation specific fashion. Next, V1 represents the scene based on Gabor filtering with intensity variations that are directed along specific orientation bands. MT pools the energy from V1 cells across space and scale and provides feedback to the V1 ventral representation.

$$f_{lgn}(x,y) = \frac{1}{2\pi\sigma_c^2}\exp\{\frac{-(x^2+y^2)}{2\sigma_c^2}\} - \frac{1}{2\pi\sigma_s^2}\exp\{\frac{-(x^2+y^2)}{2\sigma_s^2}\} \quad (1)$$

where σ_c and σ_s are the bandwidth (standard deviation) for the center and surround Gaussian profiles respectively. Through image convolution, signals containing spatial frequency confined to σ_c and σ_s are selected, which can be tuned for either magnocellular or parvocellular cells.

The temporal profile of LGN filters is based on a log-Gabor filter and is given by the frequency response:

$$F_{lgnT}(w) = \exp\{\frac{-\log(w/w_0)^2}{2\log(\sigma_t/w_0)^2}\} \quad (2)$$

where w_0 is the center frequency of the filter and σ_t controls the temporal bandwidth of the filter. At the level of LGN the visual input is represented by 16 bands corresponding to 4 spatial frequency bands and at 4 temporal (static, low, medium, and high temporal frequency) scales.

The second visual layer is marked V1 and consists of log-Gabor filters that are orientation selective and each correspond to a particular spatial and temporal frequency band. The definition of these cells is given by the frequency response:

$$F_{V1}(x,y) = \exp\{\frac{-log(x'/y')^2}{2log(\sigma_x/x')^2}\} \cdot \exp\{\frac{-y'^2}{2\sigma_y^2}\} \quad (3)$$

where $x' = x\cos(\theta) + y\sin(\theta)$, $y' = -x\sin(\theta) + y\cos(\theta)$, θ denotes the orientation of the filter, x' denotes the center frequency, and σ_x and σ_y denote the bandwidth of the filter along the x and y directions respectively. At V1, selectivity for spatial orientation is introduced and θ values correspond to an even covering in angular frequency intervals of $\pi/6$ starting from horizontal.

As M and P type V1 cells derive from the divided LGN representation, V1 is also divided into dorsal and ventral units that are distinguished by the range of spatial and temporal frequencies represented within each. The dorsal stream further includes visual area MT, that pools V1 energy from V1 dorsal units over larger receptive fields. The receptive field size at MT layer is many times the size of V1 receptive fields and pooling occurs evenly over the spatial frequency bands represented.

Feedback from MT has a multiplicative influence on ventral V1 units, so that activation among the higher dorsal MT area refines the representation that appears in V1. This interaction is multiplicative with the weighting associated with feedback connections based on the measured frequency characteristics of visual cortical cells that appear in ventral V1 and dorsal MT regions. This relationship is such that the strength of interaction corresponds to the prevalence of frequencies within ventral V1 and dorsal MT [3].

Saliency is computed on the basis of the refined ventral V1 representation. The saliency attributed to each local region of the image (represented by the vector $x_1, ..., x_n$) is given by:

$$S(x_1, ..., x_n) = -log(p(x_1, ..., x_n))$$

Saliency then corresponds to the negative log likelihood of the local feature vector computed as described in [2].

As a whole, we have shown that the model consists of a two pathway (dorsal and ventral) hierarchical representation. The computation among the higher dorsal area (MT) is used to refine the representation that appears in the early ventral area (V1v). This is consistent with the fast brain account of processing and the resulting refined ventral V1 representation is then used to determine the visual salience of each location in the image.

IV. Evaluation

In this section, we compare the model that includes recurrent feedback with the baseline algorithm to assess the extent to which the rapid recurrent processing aids in signaling targets of interest. The evaluation is aimed at determining the efficacy of the proposed approach as an attentional component of a typical machine vision system, in this case, a visual surveillance scenario. The baseline algorithm in this case consists of a feedforward model that uses ventral and dorsal representations directly to compute saliency. This is in contrast to the recurrent model, which uses the MT layer dorsal representation to refine the ventral V1 representation, with saliency then computed based on the refined ventral representation. In this way, one can determine the extent to which predictions regarding saliency benefit from recurrent refinement rather than direct use of the complete set of V1 filters. Since the baseline algorithm performs well compared to other models in the literature, it is less important how the algorithm compares with the existing body of low level local feature contrast only models of visual saliency and more important how well the addition of recurrence improves model performance.

Data employed in experimentation were collected from a number of different vantage points using a variety of cameras and varied imaging and environmental conditions. Additionally, data from public sources were used in evaluation. Qualitative evaluation was carried out on the entirety of the aforementioned data, while quantitative evaluation was performed on a representative subset of these videos for which ground truth was available or created. The ground truth for this data consisted of a set of bounding boxes for each frame of the video indicating the locations of pedestrians and vehicles in the video sequence. The intention of this labeling was to indicate targets that are not a fixed item (i.e. not part of the background). The evaluation then measures the extent to which the choice of visual representation impacts on the determination of salient targets (e.g. people and vehicles) in this context. Labeling was carried out by placing bounding boxes using the labeling software developed as part of this assessment [1]. From a qualitative perspective, one may visually inspect the saliency maps corresponding to a variety of different conditions. Figure 2 shows an example of frames sampled from videos used in the evaluation. To assess the extent to which the fast recurrent loops augment the associated determination of saliency, we consider two separate types of computation. The first computes saliency based on a log-Gabor basis oriented in the spatial dimension only (left column). The right column reflects output for log-Gabor filters that also have a temporal extent. Output is shown for the purely feedforward configuration that does not include recurrent refinement (top) and the complete model that includes such recurrence (bottom). This allows comparison of the extent to which recurrence aids in augmenting the saliency representation. As may be seen from figure 2, the recurrent feedback suppresses confidence in the saliency map attributed to spurious image structure, and affords a more complete covering of the targets of interest.

The quantitative assessment is based on two different standard metrics for assessing classifiers as follows: First a threshold is chosen to convert a saliency map to a binary classification. This is compared with the binary mask corresponding to the bounding boxes drawn for the same image. In the ideal case, the classification overlaps perfectly with the bounding boxes drawn. The nature of the classifier produced by the saliency map depends on the threshold that

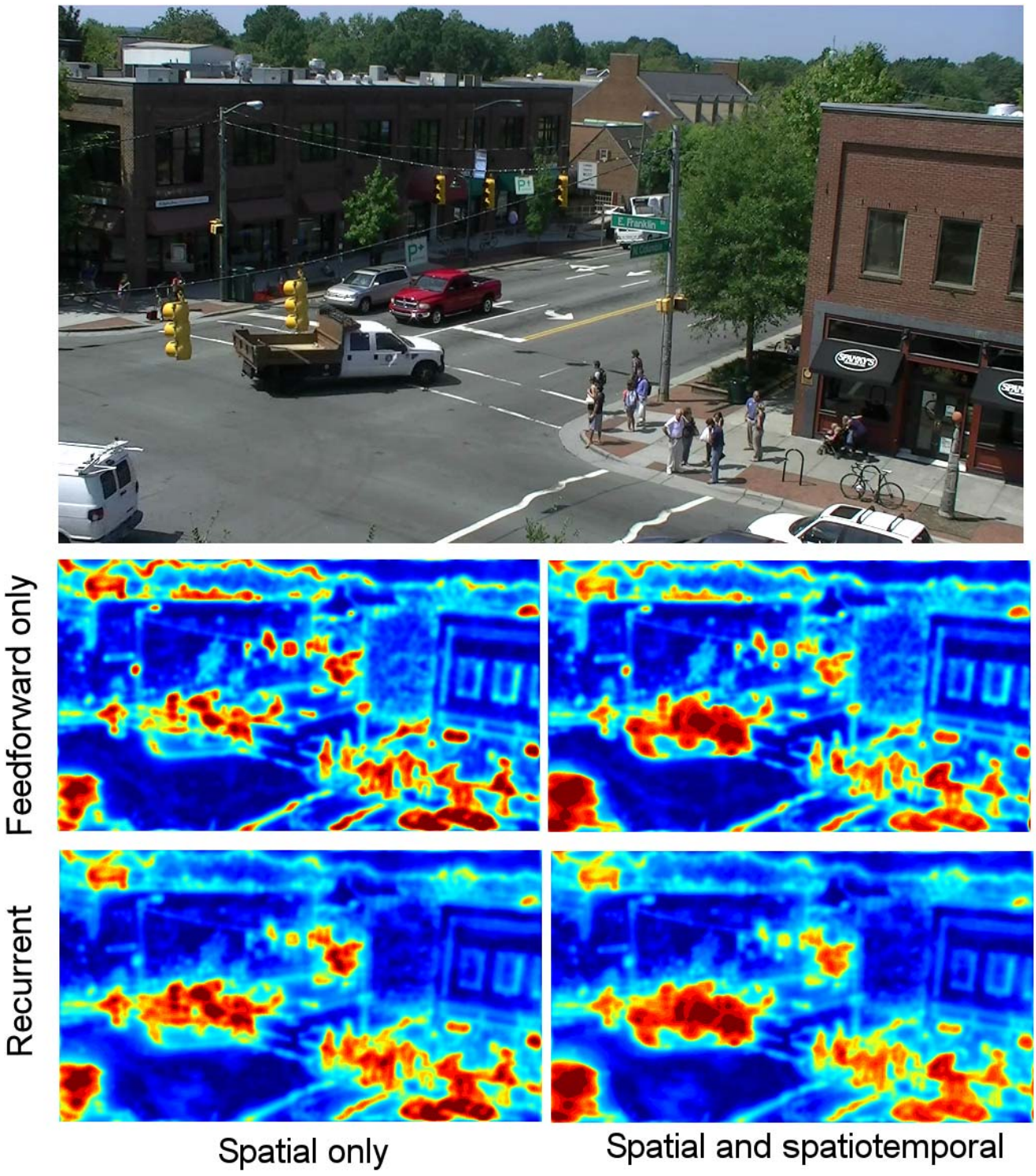

Figure 2. A quantitative comparison of classification performance in highlighting targets (foreground objects). Shown is saliency output based on a purely spatial (left two output figures) and spatiotemporal (right two output figures) basis. The two bottom figures demonstrate saliency output that includes recurrent modulation, and the output frames above demonstrates saliency with a purely feedforward approach. Note the greater emphasis on certain regions (e.g. the truck) in the case of recurrent modulation, as well as the suppression (e.g. the tree line) in the recurrent case. Hotter areas (those closer to red) in the saliency map signifies regions that are salient).

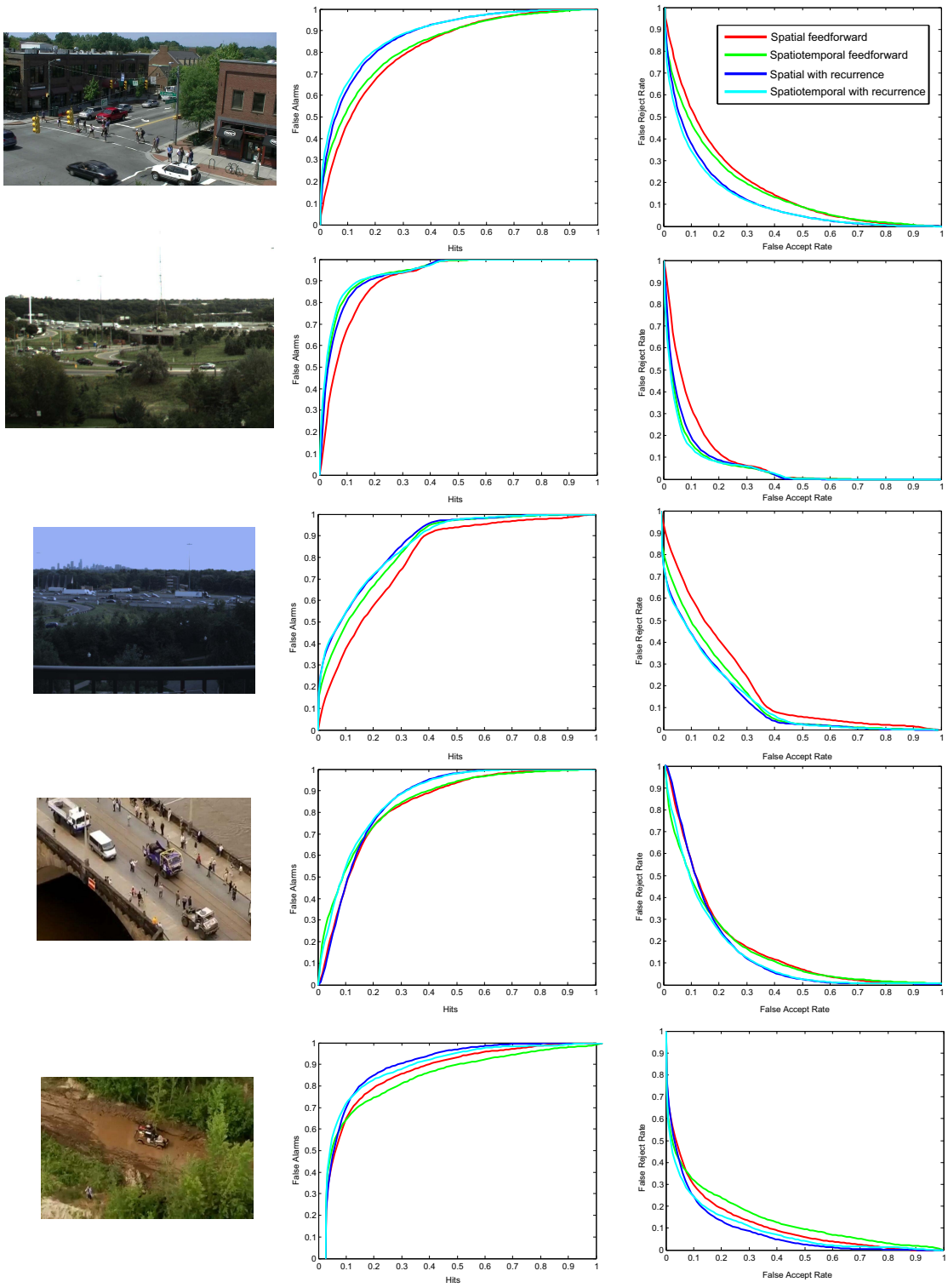

Figure 3. A quantitative comparison of classification performance in highlighting targets (foreground objects). Left: A sample frame from the video sequence characterized by the curves shown. Middle: ROC performance curves associated with the various video sequences. Right: DET curves associated with videos appearing in the data set. In general, a set of basis filters that has a temporal extent (green), outperforms a purely spatial basis set. However, output based on a processing model that includes recurrent feedback (dark blue, light blue) is even more effective in correctly predicting salient targets and augmenting the model's performance.

is chosen. In choosing a large number of thresholds from 0 to 1, an entire performance curve may be drawn for each of the methods under consideration. The specific thresholds chosen are based on the $1st, 2nd, ..., 99th$ percentile values in the saliency map. The correspondence between the classification map and the bounding box map is carried out according to 2 separate metrics as follows.

A. *ROC-curve*

The ROC curve is constructed based on analysis that is done on a pixel by pixel basis. Given a particular threshold, pixels in the saliency map are set to a value of 0 or 1 (above or below threshold). The bounding box image also specifies a value of 0 or 1 for each pixel location based on its human labeled ground truth. Four different outcomes are defined as follows: TN: True negative, 0 in saliency map and 0 in bounding box map. FN: False negative: 0 in saliency map and 1 in bounding box map. TP: True positive: 1 in saliency map and 1 in bounding box map. FP: False positive: 1 in saliency map and 0 in bounding box map. Each threshold yields a set of 4 values given by these quantities. The ROC curve depicts the true positive rate versus the false positive rate. The two extremes correspond to a 0% TP and 0% FP rate and to a 100% TP and 100% FP rate. Choosing a variety of thresholds results in a smooth curve between these two extremes.

B. *Detection Error Tradeoff*

The detection error tradeoff curve is in the same spirit as the ROC-curve. However, the precise quantity that is on display is qualitatively different. The DET curve demonstrates the number of misses (i.e. the false reject rate) versus the false accept rate. The DET curve gives a sense of how many non-target pixels are accepted versus how many on target pixels are missed. In visual surveillance scenarios, the DET curve can be an important visualization of the data since the case of false rejects can be especially important. In this scenario the DET curve then gives a sense of how many non-target pixels are accepted while maintaining a near 0 false accept rate.

Figure 3 demonstrates the ROC curves (left) and DET curves (right) associated with 4 different modes of computation. These four cases arise from the inclusion/exclusion of cells that have a spatiotemporal (as opposed to purely spatial) extent, and from the inclusion/exclusion of recurrent refinement. The output corresponding to the labeling of 5 different video sequences is considered, with bounding boxes associated with all of the persons/vehicles in the scene. As may be seen, the results are quite consistent across the entire range of videos considered. In the feedforward case, the inclusion of units that have a temporal extent (green) is revealed to be important in the overall representation of saliency as compared with the filters that have a purely spatial support (red). Importantly, the saliency tied to the representation that includes recurrent refinement shows improved performance due to greater emphasis on targets of interest and better suppression of structured background elements that do not conform to content of interest in this scenario. This is true of both the spatial only (dark blue) and spatiotemporal (light blue) cases included in the evaluation.

One question that arises concerning evaluation, is that of whether bounding boxes suffice to adequately represent performance differences (as they include background pixels), as opposed to very detailed masks that conform to precise target boundaries. To this end, we performed an additional test to verify that conclusions drawn from a bounding box based assessment are in agreement with an evaluation based on precise masks. This evaluation was carried out on 12 frames of a video sequence with a seed frame chosen at random and the remaining frames spaced at intervals of 500 frames from the seed frame and each other. For each of these 12 frames, a precise ground truth mask was manually drawn. An example of the precise mask corresponding to one of these frames appears in figure 4.

Figure 4. An example of the scene labeling used in the quantitative performance evaluation.

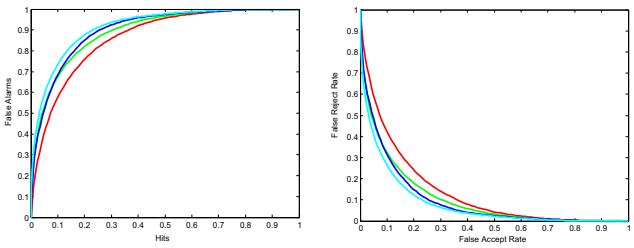

Figure 5. Performance evaluation based on the fine masks for the data set shown in figure 2 (Top row).

As may be seen from the results in figure 5 (compare with figure 3 top row), the quantitative assessment based on a precise mask is consistent with expectations based on the evaluation that uses more precise masks that capture the precise form factor of targets.

While there exists a small difference, as expected, in the absolute classification performance associated with the algorithms tested, the ranking and distance between the curves remains very similar. This provides a reasonable level of confidence that the results attributed to the bounding box based masks is predictive of performance that is gauged on the precise localization of salient targets.

It is evident from the evaluation that the recurrent modulation from *dorsal type* features has an appreciable impact on model behavior. In particular, the mechanism appears to place more emphasis on targets of interest, while suppressing saliency attributed to spurious background elements. In addition, there is improvement in the precise localization of target items and they are also better separated. An additional benefit is the fact that the suppression of spurious elements in the saliency map implies less *hot spots* in the resulting saliency map. In the context of an attention/recognition system, this implies fewer regions require consideration by the recognition system, improving the efficiency of the overall system.

As a whole, this establishes the importance of the inclusion of a fast recurrent mechanism of the sort that appears in a *fast brain* theory of recurrent visual representation and computation. This implies an important role for a heretofore ignored forest-trees mechanism that is independent of any recognition scheme or semantic labeling.

V. Discussion

In this paper, we have put forth a novel mechanism for attentive processing inspired by the *fast brain* theory appearing in the biological vision literature. Not unlike the *gist* based account of scene representation, this strategy presents an alternative novel means of allowing a rapid more global sampling of the scene to impact on more localized and specific processing.

It is revealed that this strategy, based on rapid recurrence, provides substantial gains in attributing saliency to targets of interest in a typical surveillance and recognition scenario. It is also demonstrated that a strategy that uses rapid recurrence from a dorsal stream representation to refine the ventral stream representation, outperforms the strategy of computation that considers only feedforward computation using the ventral and dorsal representations directly, and without recurrent refinement.

Although the focus in the discussion appearing in this paper is on visual salience, this proposal presents an attentive process that may be separated from the determination of saliency as the resulting refined representation might be employed for any manner of visual processing task.

Unlike the gist based proposal, that posits fast recognition of scene properties to drive location based spatial bias, the proposal at hand implements an implicit bias in both space and frequency with broader receptive fields providing a coarse sampling of scene content that guides further processing. Importantly, this is done in a manner that does not require any explicit training of prior contextual models, or ties to semantic labeling.

Acknowledgment

The authors gratefully acknowledge the support of NSERC in funding this work. John K. Tsotsos is the NSERC Canada Research Chair in Computational Vision.

References

[1] Anonymous. Anonymous software tool, 2011. Ground truth labeling software.

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

[3] J. Bullier. Integrated model of visual processing. *Brain Res. Brain Research Reviews*, 36(2-3):96–107, 2001.

[4] M. Greene and A. Oliva. The briefest of glances: the time course of natural scene understanding. *Psychological Science*, 20(4):464–472, 2009.

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

[6] D. Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9:353–383, 1977.

[7] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 2002.

[8] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39:3157–3163–778, 1999.

[9] A. Torralba, A. Oliva, M. Castelhano, and J. M. Henderson. Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, 2006.

[10] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[11] J. K. Tsotsos. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2).

[12] J. K. Tsotsos. Analyzing vision at the complexity level. *Behav. Brain Sci*, 13:423–445, 1990.

[13] J. Xu, Z. Yang, and J. Z. Tsien. Emergence of visual saliency from natural scenes via context-mediated probability distributions coding. *PLoS ONE*, 5(12):e15796. doi:10.1371/journal.pone.0015796, 2010.