Selective Visual Attention in Electronic Video Surveillance

James Mountstephens, Craig Bennett and Khurshid Ahmad

Department of Computing, University of Surrey

Abstract. In this paper we describe how a model of selective visual attention, driven entirely by visual features might be used to attend to "unusual" events in a complex surveillance environment. For the purposes of illustration and elaboration we have used an implementation of an early processing model of attention (due to Itti and Koch [1]) to process ground-truth surveillance video data [2].

1 Introduction

Human beings have learnt to, and are genetically capable of, focussing on certain features of their visual environment to the relative exclusion of other features. Psychologists, neurobiologists and latterly surveillance experts and computing professionals ask how it is that human beings can selectively attend to some features of the environment in the light of the observation that there is no fixed correspondence between stimuli, their properties, and attention objects [3]. There are speculations that experience, strategy and/or individual capabilities may influence the connection between elementary stimulus units, thereby affecting the ways they combine to create a single *complex attention object* [ibid; our emphasis].

A single complex attention object is critical for electronic video surveillance, an active area of research drawing on work in computing, psychology and security. The purpose of a surveillance system is to detect unusual events occurring in the surveilled area based on visual input and perform some action in response. Unusual events might include fighting or abandoning bags [4] and the required response might be to sound an alarm [5]. By understanding the cues leading to unusual events, an expert surveillance operative knows what to look for and what to track in a visual scene. It is this level of performance that electronic surveillance systems attempt to emulate but although some widely-used techniques exist [6, 7], no general machine solution is currently available. The basic property of attention is that of *selectivity* and the mode of selectivity most relevant to surveillance is that attention can "direct our gaze towards objects of interest in our visual environment" [8, pp1]. Before higher-level tasks such as pattern recognition, scene understanding or description [9] can take place we must know *where* to look and part of our aim is to investigate the role that this attentional guidance can play as part of a larger architecture designed for visual understanding [10, 11].

Mountstephens J., Bennett C. and Ahmad K. (2005).
Selective Visual Attention in Electronic Video Surveillance.
In *Proceedings of the 5th International Workshop on Pattern Recognition in Information Systems*, pages 198-203
Copyright © SciTePress

In this paper we describe how a model of visual attention, driven entirely by visual features might be used to attend to "unusual" events in a complex environment. We have used an *early processing* model of visual attention which has been implemented by Itti and Koch [1, 12]. The images used in our experiment were created by the CAVIAR project which was focused, in part, on public space surveillance task, and 'are ground truth labeled frame-by-frame with bounding boxes and also a semantic description of the activity in each frame' [2, pp.1].

2 A Method for Computing Salience and Conspicuity

There are several computational models of visual attention [eg. 13, 14] and among them there is a distinction between *bottom-up* and *top-down* processing. Bottom-up processing is reactive, based entirely on features of visual input whereas top-down processing allows prior knowledge to bias response to input in order to intentionally promote or suppress certain features of it. We have selected an implemented and freely available model of visual attention due to Itti and Koch [15] which in its basic form is purely bottom-up (though see [16] where goal-orientated attentional guidance is explored). This model is based on earlier work by Koch and Ullman [17] which itself drew heavily on the *feature integration* model of attention by Treisman and Gelade [18]. The input to the model is a video sequence of *n*-image frames. There are two outputs of the model which are important to us, namely that its output is a single location in the image frame (a gaze *fixation*) and that this location will move during a frame sequence due to the mechanism of *inhibition of return.* These outputs are the product of a number of intermediate processes in which image features are both *extracted* and *combined* to yield a location that is currently the focus of attention, or is *salient* in the parlance of Itti and Koch.

A *saliency map* is calculated to encode the conspicuity of every point in the input image and consists of a weighted sum of early visual features calculated in parallel at every location. The most common features, and those used in our experiments, are colour, intensity, orientation (0, 45, 90, 135 degrees), movement (left, right, up, down) and flicker. Each raw feature map is calculated at multiple spatial scales (in a Gaussian pyramid) and differencing across these scales is used to calculate a *centre-surround* response at every location so that local contrast rather than the absolute value of the feature becomes important. A process of normalisation and combination across scales yields a single *conspicuity map* for each feature and the saliency map is constructed as a weighted sum of these conspicuity maps. The saliency map is realised as an array of leaky integrate-and-fire neurons and a winner-takes-all process of competition is performed to locate the most salient point in the map. This winning point is the output of the model and is the suggested location for the gaze. However, given a static image and constant model parameters, the winning location will always be the same so finally a mechanism for *inhibition of return* operates to enable the image to be scanned in order of decreasing salience. By giving the saliency map a large negative weighting in the region centred on the current fixation, the fixation point must move to another location at the next time step. Strategies for combining conspicuity maps to build a saliency map are discussed in [19]. Figure 1 comprises a pseudo-code for calculating the current fixation location.

```
<CALCULATE CURRENT FIXATION LOCATION>
 →[video sequence of n image frames]
  For every frame in the video sequence(0..n)
    For all features to be extracted (intensity, colour, orientation, flicker, motionL/R/U/D)
      <construct 8-spatial scale Gaussian pyramid of feature>
       For all spatial scales(i..8)
         <extract raw image feature at spatial scale i>
         →[Raw image features]
       end for
       → [Gaussian Pyramids of Raw image features]
     <combine raw features across scales in pyramid and normalise>
     → [Normalised feature maps]
     For all feature maps (1..2)
       <combine normalised feature maps across scales and normalise>
       → [Normalised conspicuity maps]
     end for
    end for
    <weighted summation of  conspicuity maps>
    → [Saliency map]
    <winner-takes-all search of saliency map>
    → [Current most salient location]
    <inhibition of return>
    → [Inhibited saliency map]
  end for
end process

                    Key:  Control  → [Data]   <process name>
```

Fig. 1. Pseudocode description of Itti and Koch attentional model. Computations of static image features of colour, intensity and orientation features can be found in [20] and that of the dynamic features of flicker and motion are in [21].

Given that the output of the preceding model is a location in an image, we asked whether this calculated location would correspond to locations worth attending to in a surveillance situation. As suggested earlier, this gaze location guidance is likely to be a prerequisite for higher-level visual processing. In order to perform some elementary experiments addressing this question we employed two main practical components: a software implementation of the Itti and Koch model (iNVT - the iLab Neuromorphic Vision Toolkit [12]) and a publicly-available surveillance video dataset (CAVIAR dataset [22]). The iNVT accepts images or appropriately named image sequences as input and can produce output images showing the trajectory of fixation points. The CAVIAR dataset comprises eighty video clips (over 90,000 frames) containing six event-types: *walking*, *browsing*, *collapsing*, *leaving object*, *meeting*, and *fighting* [2] and each clip is accompanied by a 'semantic' description of objects of interest in the scene. In this case, objects of interest are moving people or discarded objects and per-frame information about these people, their activities and bounding box information (including centre coordinates, width and height) has been hand annotated for all clips. Of the six event-types in the CAVIAR dataset, the category of "leaving object" was selected for this experiment because it typically contains both usual and unusual events – the *usual* event of walking precedes the *unusual* event of leaving an object. In our experiment we run the five "leaving object" videos, per-frame, through the iNVT to produce a gaze fixation point and use the bounding box information in the ground-truth description accompanying those videos to decide whether the fixation

point is on an object of interest. For every frame, if the fixation point is within the bounding box of a named object (as defined in the CAVIAR ground-truth description) we construe this as a *hit*. Otherwise it is a miss. From this we can calculate both the total number of hits (on *any* person or object in the scene) and the total hits on the object associated with the unusual event (the *discarded* object). With this hit data and knowledge of the total number of frames in the video we can calculate some elementary performance ratios.

3 Experiment: Attending to Discarded Objects

Using the method described above, five CAVIAR videos of people walking then discarding objects were processed by the iNVT. Table 1 shows the percentage of fixation point hits for all objects in the scene, listed in descending order of hit rate for the five videos processed, and whether the event of discarding an object was hit.

Table 1: Total object hit rate and event of discarding object detection

Video	Hit %	Event of Discarding Object Hit?
LeftBox	40	No
LeftBag_PickedUp	38	Yes
LeftBag	37	No
LeftBag_AtChair	33	No
LeftBag_BehindChair	30	No

Considering that no explicit knowledge is embodied in the attentional model, column 2 suggests an encouraging degree of correspondence between the features driving the model and events in the scene. Of these hits it makes sense to ask how many of those locations were important to the task, namely how many correspond to *unusual* events. Column 3 shows whether the event of *leaving an object* was hit by the fixation point and here, the results are less encouraging - only one event of discarding an object was hit. A visual examination of model output by us demonstrated that salience is affected by a) what has happened in the last time-step, through inhibition of return, and b) what is going on elsewhere in the scene. For example, in some cases the gaze was close to the individual who would soon discard the object but could not move to that location at the point of jettison since it had been inhibited in the saliency map. Also, the usual events of innocent people walking could divert attention away. In terms of visual features alone, the act of discarding an object was often not salient enough to attract the focus of attention. An examination of other activities in the CAVIAR dataset has produced more encouraging results and will be reported in due course.

4 Conclusions and Future Work

Itti and Koch's model produces an annotation of an image by marking changes in the focus of attention by encircling objects 'behaving' unusually in the image and by

marking the last focal point. The model identifies some of the focal objects correctly with the limited information it has. Our future work will include modulating the Itti and Koch model's mechanism for inhibition of return to improve event tracking performance and to investigate how these salient events can be learned automatically.

We are exploring how the annotation produced by the model can be used as a part of an overall annotation framework where advantage is taken of other modalities used in describing objects in the image. Elsewhere, we have used collateral linguistic description of still images for the purpose of indexing and retrieving scene of crime images [23]. We have begun to obtain a collateral linguistic description of a small subset of the CAVIAR video clips to obtain a more meaning-oriented information about objects in the videos; surveillance experts are being interviewed for this purpose. The longer-term aim is to combine visual attention analysis with information obtained from other modalities used to describe the same environment, specifically a linguistic description.

This research is currently at an early stage and forms part of the EPSRC (Engineering and Physical Sciences Research Council) funded project REVEAL (Recovering Evidence from Video). The strategic objective of the project is to promote those key technologies which will enable automated extraction of evidence from CCTV archives. In addition to the development of methods for capturing the conceptual structure underpinning the work of surveillance experts, the project aims to develop methods of integrating the linguistic structure (the Visual Evidence Thesaurus) and the visual-content (the Surveillance Meta-Data Model) through co-learning to enable the automatic annotation of video data streams, and facilitate the retrieval of video evidence from high-level queries. [24].

Acknowledgements

This work has been supported by the UK Engineering and Physical Sciences Council's grant REVEAL, (GR/S98443/01). The project is being conducted in close collaboration with Kingston University and Sira Ltd, and supported by Police Information Technology Organisation (PITO) and Police Scientific Development Branch (PSDB). James Mountstephens is funded through an EPSRC Studentship and Craig Bennett funded by a University of Surrey Studentship.

References

1. Itti, L. and Koch, C. (2001), "Computational Modelling of Visual Attention", Nature Reviews Neuroscience, Vol. 2(3), pp 194 – 203.
2. Fisher, R. (2004). "The PETS04 Surveillance Ground-Truth Data Sets", Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance, pp 1-5.
3. Eysenck, M. W. (Ed.) (1990), The Blackwell Dictionary of Cognitive Psychology. Oxford : Blackwell Reference, 1990.
4. Haritaoglu, I., Harwood, D. and Davis, L. S. (2000), "W4: Real-Time Surveillance of People and their Activities", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22(8), pp 809 – 830.

5. Foresti, G. L., Marcenaro, L., and Regazzoni, C. S. (2002), "Automatic Detection and Indexing of Video-Event Shots for Surveillance Applications", IEEE Transactions on Multimedia, Vol. 4(4), pp 459 – 471.
6. Aggarwal, J., Cai, Q. (1997), "Human Motion Analysis: a Review". Proc. IEEE Nonrigid and Articulated Motion Workshop, pp 90 – 102.
7. Gavrila, D. (1999), "The Visual Analysis of Human Movement: a Survey", Vision and Image Understanding, Vol. 73(1), pp 82 – 98.
8. Itti, L. (2003), "Visual Attention", In M. A Arbib,. (Ed), The Handbook of Brain Theory and Neural Networks, 2nd Ed. MIT Press, pp. 1196-1201.
9. Rittscher, J., Blake, A., Hoogs, A. and Stein, G., (2003), "Mathematical Modelling of Animate and Intentional Motion", Philosophical Transactions: Biological Sciences. Vol. 358(1431), pp 475 -490
10. Marr, D. (1980), "Visual Information Processing: the Structure and Creation of Visual Representations". Philosophical Transactions of the Royal Society of London B, 290: pp. 199 – 218.
11. Ullman, S. (1984), "Visual Routines", Cognition, Vol. 18, pp 97 – 159.
12. http://ilab.usc.edu/bu. Last accessed 17-03-05.
13. Wolfe, J. (1998), "Visual Search: a Review". Attention, H. Pashler (Ed.), London UK: University College Press.
14. Tsotsos, J. K. , Culhane, S.M., Wai, W. Y. K., Lai, Y. H., Davis, N. & Nuflo, F. (1995), "Modelling Visual-Attention via Selective Tuning", Artificial Intelligence, Vol. 78 (1-2), pp 507-45.
15. Itti, L., Koch, C., Niebur, E. (1998), "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20(11), pp 1254-1259.
16. Navalpakkam, V., and Itti, L. (2002), "A Goal Oriented Attention Guidance Model", Lecture Notes in Computer Science, Vol. 2525, pp. 453-461.
17. Koch, C, Ullman, S. (1985), "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry". Hum Neurobiol, Vol 4(4), pp 219 – 227.
18. Treisman, A. M., Gelade, G. (1980), "A Feature-Integration Theory of Attention", Cognit Psychol, Vol. 12(1), 97-136.
19. Itti, L., Koch, C. (2001), "Feature Combination Strategies for Saliency-Based Visual Attention Systems", Journal of Electronic Imaging, Vol. 10(1), pp. 161-169.
20. Itti, L. (2000), "Models of Bottom-Up and Top-Down Visual Attention", PhD thesis, California Institute of Technology.
21. Itti, L. Dhavale, N. Pighin, F. (2003), "Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention", Proc. SPIE 48th Annual International Symposium on Optical Science and Technology, pp. 64-78.
22. http://www.dai.ed.ac.uk/homes/rbf/CAVIAR/. Last accessed 17-03-05.
23. Ahmad, K., Tariq, M., Vrusias, B. and Handy C. (2003), "Corpus-Based Thesaurus Construction for Image Retrieval in Specialist Domains". In (Ed). Fabrizio Sebastiani. Proc 25th European Conf on Inf. Retrieval Research (ECIR-03, Pisa, Italy) LNCS-2633. Heidelberg: Springer Verlag. pp 502-510.
24. www.computing.surrey.ac.uk/ai/reveal/. Last accessed 17-03-05.