VISUAL SALIENCE AND PRIORITY ESTIMATION FOR LOCOMOTION USING A DEEP CONVOLUTIONAL NEURAL NETWORK

N. Anantrasirichai, Iain D. Gilchrist and David R. Bull

Bristol Vision Institute, University of Bristol, Bristol BS8 1UB, UK

ABSTRACT

This paper presents a novel method of salience and priority estimation for the human visual system during locomotion. This visual information contains dynamic content derived from a moving viewpoint. The priority map, ranking key areas on the image, is created from probabilities of gaze fixations, merged from bottom-up features and top-down control on the locomotion. Two deep convolutional neural networks (CNNs), inspired by models of the primate visual system, are employed to capture local salience features and compute probabilities. The first network operates through the foveal and peripheral areas around the eye positions. The second network obtains the importance of fixated points that have long durations or multiple visits, of which such areas need more times to process or to recheck to ensure smooth locomotion. The results show that our proposed method outperforms the state-of-the-art by up to 30 %, computed from average of four well known metrics for saliency estimation.

***Index Terms*—** Salience, convolutional neural network, deep learning, locomotion

1. INTRODUCTION

As we perceive a huge amount of visual information continuously, our nervous system makes decisions on which parts of this information should be further processed, and also prioritises it. Similarly, intelligent machines are required to distinguish and prioritise important information. This process is not straightforward and it is even more complicated when visual information is obtained during locomotion, because the dynamic scene we see moves towards us and hence changes continuously. This visual information is required to obtain a sense of distance, global information about self-motion through an environment and the posture of the body relative to the environment. Achieving prioritisation of visual information will improve decision performance for autonomous robots and guidance aids for the visually impaired.

Saliency-based modelling is one of the most successful approaches to this problem for static images [1]. However, this technique was not developed for visual information obtained during locomotion, and it is not directly transferable to this scenario. In this work we specify the context of the video as resulting from locomotion and therefore model a *priority map* reflecting the combined representations of salience (bottom-up) and relevance (top-down) in the selection process, which best describe the firing properties of neurons [2]. Salience is the low-level property of the scene, where it is prominent compared to its neighbours, whilst relevance exploits top-down factors, e.g. expectation and experience, to determine attentional allocation.

In this paper, we exploit visual information from human eye fixations during walking on complex terrains. Eye positions are acquired with a mobile eye tracker. The areas around the eye positions are employed to train two convolutional neural networks (CNN). The first CNN captures instantaneous local visual features perceived during locomotion. This replicates human visual system by using centre-surround inputs, inspired by neural responses in the lateral geniculate nucleus (LGN) [3]. The second CNN determines which areas on the terrain are fixated upon for a long duration or are the subject of repeated fixations, as these are likely to relate to the significance of these areas. The priority map is finally created from merged probabilities of gaze fixations from both networks.

The remaining part of this paper is organised as follows. Section 2 presents background and related work. Then, we describe our proposed method of priority estimation in Section 3 and the results are shown in Section 4. Finally the conclusions and future work are set out in Section 5.

2. BACKGROUND AND RELATED WORK

In early work, prioritising visual information was performed through the properties exhibited in the scene, following the assumption that attention is influenced by stimuli. Bottom-up methods have hence been developed by replicating the early processes of the primate visual system [3], where specific types of neurons detect features of stimuli, such as edges and noisy background in the natural environment.

Itti et. al. [4] modelled visual attention based on the feature integration theory, where elementary features, e.g. colour, intensity and orientations are represented in the visual cortex. Multiple scales were employed through a set of weighted centre-surround outputs in [5]. An intensive survey of methods mimicking the human visual system can be found in [6]. Statistical models and image processing techniques

are also employed. Zhang et al. modelled a Bayesian framework from the self-information of visual features, and overall saliency emerged as the point-wise mutual information [7]. In [8], saliency was determined by quantifying the joint likelihood and self-information of each location image patch. Hou et.al. [9] introduced a saliency map based on an image signature that spatially approximated the foreground of an image and predicted fixation points using a Discrete Cosine Transform.

Three dimensional data has also been employed as humans perceive visual information based both on the current scene (spatial information) and on the accumulated knowledge from previous frames (temporal information). Apart from fundamental features, motion was captured by applying optical flow [10], or three dimensional textures [11]. All of these methods aim to detect moving objects in a static scene or a slow panning background.

Later top-down mechanisms were included in the process based on experiments that showed the relationship between bottom-up and top-down attentions employed to process complex dynamic content [12]. Judd et. al. [13] employ three levels of features, i.e. i) low-level physiologically plausible features, ii) mid-level features such as the objects at the horizon, and iii) high-level features such as people and faces. Object recognition is used in [14] and task-driven object based model by a Bayesian framework is developed in [15] to deal with complex environments.

3. PROPOSED SCHEME

3.1. Training by two parallel CNNs

A CNN is a biologically-inspired architecture that comprises multiple layers of neuron collections that have learnable weights and biases. Their results are tiled so that they overlap to obtain a better representation of the original image. The CNN creates its filters' values based on the task. Generally the CNN learns to detect edges from the raw pixels in the first layer, then uses the edges to detect simple shapes in the next layer. The higher layers produce higher-level features.

The proposed framework is shown in Fig. 1, where two independent CNNs are employed to compute the probability of being a fixation (Section 3.1.1) and the probability of the fixation having a long visit or multiple visits (Section 3.1.2). We develop our network using the Caffe framework with parameters recommended in Alex Krizhevsky's ConvNet [16]. The network consists of three layers of convolution, max-pooling, rectified linear unit (ReLU), and local normalisation, followed by a fully connected layer and a linear classifier at the top. Deeper networks may be used which generally give better performance. The three-layer network is employed here because of the limitations of our computational system.

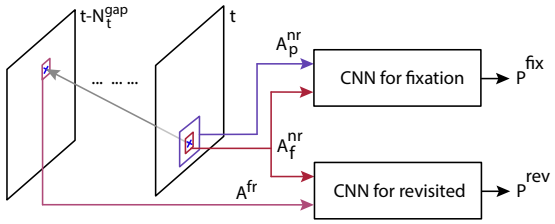

Fig. 1. The proposed framework using two parallel CNNs to estimate probability of being a fixation and a revisited fixation.

3.1.1. Fixation probability estimation

The proposed process of probability estimation is illustrated in Fig 2. Two classes are labelled as fixations (acquired by mobile eye tracker) and random points (selected randomly from the same distribution of fixations). For each point, two different sized areas centred at it are used, namely fovea ($h_f \times h_f$ pixels) and peripheral areas ($2h_f \times 2h_f$ pixels). The peripheral area is resized to the fovea size, $h_f \times h_f$ pixels. Then, these two areas, 6 colour channels in total, are combined into one three-dimensional input to train the CNN. Both areas are employed following the idea of centre-surround processes, inspired by neural responses in the lateral geniculate nucleus (LGN) [3]. Note that fusions of these two streams at the intermediate layers or at the last layer may worsen the network performance by up to 3 % in term of classification accuracy, since the relationship between fovea and peripheral areas is not used in the early stage as it should be. The output of the network is the probability of being a fixation P^{fix}.

3.1.2. Long-duration or revisited-fixation model

The second CNN is employed to distinguish the areas that are fixated upon for longer than others or the areas that are fixated upon twice or more. Two classes are defined in the training process, namely i) short and single-visit fixations and ii) long or revisited fixations. The short and single-visit fixations represent the eye position appearing once in the whole sequence, whilst long or revisited fixations represent when the subject looks at the particular area, then tracks for a number of frames or saccade away and come back to fixate at it later on.

We estimate the probability P^{rev} of the long or revisited fixation using contents from areas both near and far away. The near area A_t^{nr} and its corresponding far area A_t^{fr} are the equivalent fovea-sized regions around the eye positions at the current frame t and the frame $t - N_t^{gap}$, respectively, where N_t^{gap} is the number of frames between the current frame and its furthest previous frame, where A_t^{fr} exists. Here, A_t^{fr} is searched within 5-20 frames[1], $N_t^{gap} \in [5, 20]$, and optical

[1] Based on the average walking speed of 5 km/h and our SMI eye tracker's specifications [17], the area where the participant looks will no longer be in the frame 20 frames later.

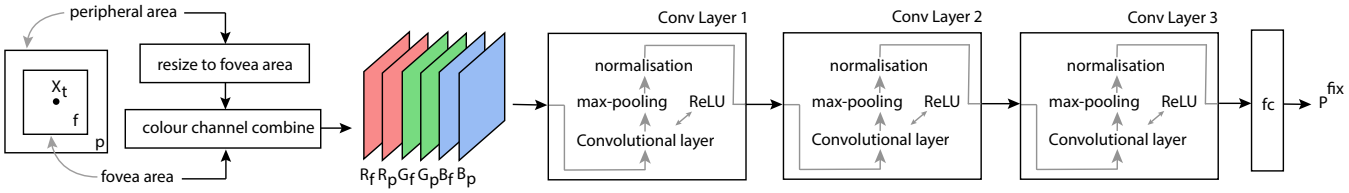

Fig. 2. The proposed method of probability estimation of fixations for locomotion.

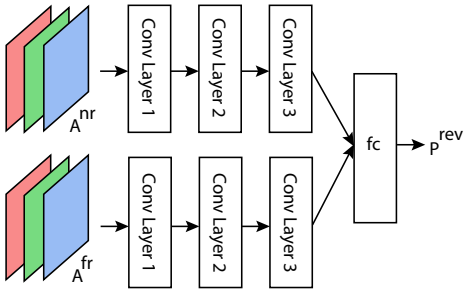

Fig. 3. The process of revisited fixation model estimation using the near area $A_t^{nr} = \{R_f^{nr}, G_f^{nr}, B_f^{nr}\}_t$ and the far area $A_t^{fr} = \{R_f^{fr}, G_f^{fr}, B_f^{fr}\}_t$

Fig. 4. Eye tracking sequences containing a variety of ground materials. The circles show fixated points.

flow is estimated using the RANSAC technique [18]. Subsequently, A_t^{nr} and A_t^{fr} are fed into two-stream CNN that joins at a fully connected layer as shown in Fig. 3. This network gives better classification accuracy over the one-stream CNN (joined inputs similar to Fig. 2) by up to 4 %. This structure is similar to the human visual process, where the low-level features of the images acquired in different times are extracted separately.

3.2. Priority map construction

A priority map S_k is constructed using the models generated following Section 3.1.1 and Section 3.1.2. To reduce computational time, only the key points are processed. Our previous work in [19] found that humans search for safe foot placement and they also observe the edges of a path as a guide for safe traversal of the terrain. Therefore, we first segment each frame into non-overlapping regions [20] and the key points are the middle points in each regions and 100 random points located on the boundaries between regions. The area A_k^{nr} and peripheral area around the key point k are employed to compute P_k^{fix}. If the corresponding area A_k^{fr} on the previous frame $t - N_t^{gap}$ of the key point k is found, P_k^{rev} is also computed. Similar to training process, A_k^{fr} is searched within 5-20 frames, $N_t^{gap} \in [5, 20]$. The final probability is

$$P_k = \max(P_k^{fix}, P_k^{rev}) \quad (1)$$

S_k is interpolated to the frame resolution using bicubic interpolation. Finally, as the eye positions exhibit centre-bias behaviour (head is often moved to improve vision), we simply apply a Gaussian weight with $\sigma = |\frac{\min(W,H)}{4}|$ to S_k, where W and H are the width and the height of video resolution.

4. RESULTS AND DISCUSSION

The test sequences were acquired using the SensoMotoric Instruments (SMI) Eye Tracking Glasses. These produce a point of view video at a resolution of 1280×960 pixels ($W \times H$) at 30 fps. The system provides a scene field of view of 60° horizontally and 46° vertically. So, the fovea and peripheral areas used in our method are approximately 64×64 pixels ($h_f = 64$) and 128×128 pixels, respectively. Only fixations were used. Saccades and noise were removed using [21].

We tested the proposed scheme using 6 sequences with eye tracking from 6 participants walking on various sloped terrains containing mixed materials of dirt, rocks, grass and woods as shown in Fig. 4, and they vary between approximately 4-6 minutes in duration. To ensure that the random points were sampled from the distribution of human eye fixations as recommended in [22], their locations were randomly selected from one of the fixation points of all training sequences. The objective results were evaluated using i) normalized scanpath saliency (NSS), ii) linear correlation (CC), iii) Area Under ROC curve measure based on Ali Borji's method (AUC-Borji) [23], and iv) Area Under ROC curve measure based on Judd's method (AUC-Judd) [24].

4.1. Proposed model testing

We first investigated the performance of the proposed method for individual participants. A 2-fold cross validation was em-

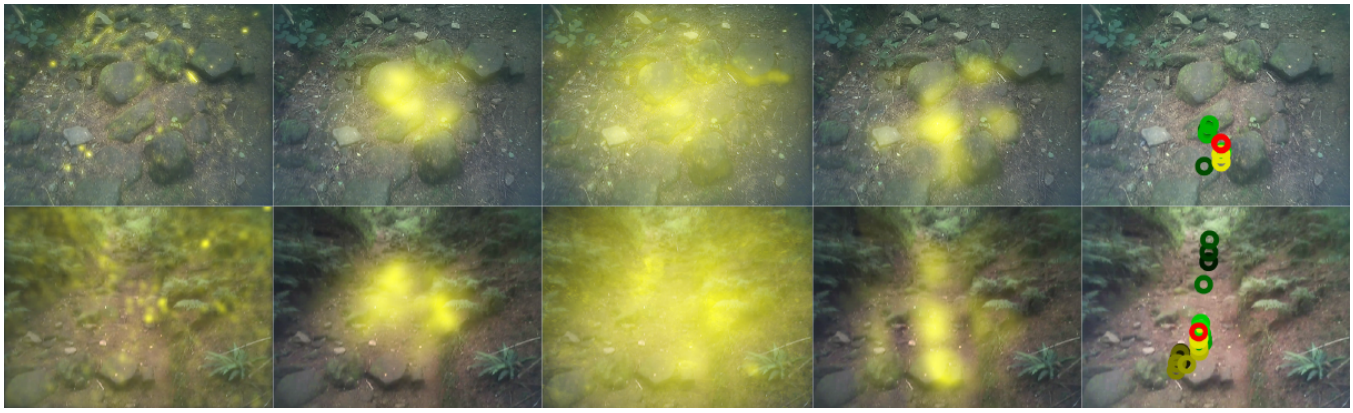

image signature CovSal Judd Proposed eye tracking

Fig. 5. Priority maps (brighter yellow is higher priority) generated using i) image signature [9], ii) region covariances (CovSal) [25], iii) multi-level features (Judd) [13] and iv) our proposed method. Right images show eye position at the current frame (red) and warped eye positions from the previous frames (green) and the future frames (yellow) - darker is further from the current frame.

Table 1. Performance of the proposed method on the individual participant measured by the average between AUC-Borji and AUC-Judd

model	#1	#2	#3	#4	#5	#6	mean±std
P_k^{fix}	0.94	0.88	0.94	0.82	0.88	0.90	0.89±0.047
P_k^{rev}	0.90	0.88	0.89	0.87	0.86	0.89	0.88±0.014
P_k	0.96	0.93	0.96	0.89	0.92	0.94	0.93±0.027

Table 2. Performance comparison using several metrics for saliency assessment

method	NSS	CC($\times 10^{-3}$)	AUD-Borji	AUD-Judd
Signature [9]	0.36	0.36	0.56	0.77
Judd [13]	1.21	1.22	0.82	0.90
CovSal [25]	1.36	1.67	0.77	0.90
proposed P_k^{fix}	1.80	1.89	0.88	0.93
proposed P_k	**2.01**	**2.03**	**0.91**	**0.95**

ployed - the first half of the sequence was used for training and the second half was used for evaluation. Then, they were swapped and the results were averaged. Table 1 shows the average of the areas under the ROC curve computed by Ali Borji's method and Judd's method. The means and the standard deviations (last column) show that the proposed method achieves consistent performance for individual participants as the standard deviation is not high. Using two parallel CNNs can improve the system performance by up to 5.5 %.

4.2. Performance comparison

Here we used 2-fold cross validation (3 sequences were used for training and the other 3 sequences were used for testing) resulting in 20 cross-validation tests in total and the results were averaged. We compared our results to those of i) image signature [9], ii) multi-level features (Judd) [13], and iii) region covariances (CovSal) [25]. The objective results tabulated in Table 2 clearly show that our proposed method outperforms the state-of-the-art with the improvement of the NSS score by 48 %, the CC score by 22 %, the AUD-Borji score by 11 %, and the AUD-Judd score by 6 %. The results also show that including the information about the long and revisited fixations can improve the prediction performance by approximately 6 %.

Fig. 5 shows the priority maps overlaid on the images. In the case of complex terrain, almost everywhere in the scene has high saliency leading to the difficulty of prioritisation – we can see that the results of the image signature and Judd show bright yellow areas all over the images. Our priority maps provides the closet match to the ground truths. This could be because our local visual features (bottom-up) used for saliency estimation are extracted using the models trained by task-driven information (top-down).

5. CONCLUSIONS AND FUTURE WORK

We have presented a novel learning-based method for saliency and priority estimation of human fixations during locomotion. Two parallel convolutional neural networks are employed to extract local visual features from areas around fixated eye positions and features to identify long fixations or multiple visited fixations. The probabilities computed from both networks are merged to create a priority map, which can be used by autonomous machines or human guidance systems, to improve their decision performance. Our framework outperforms existing methods by up to 30 % measured from well-known metrics for saliency estimation. For future work, the proposed framework will be validated with more terrain types and eye movement patterns will be included in the system to improve the accuracy of fixation prediction.

6. REFERENCES

[1] Ali Borji and Laurent Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 185–207, 2013.

[2] J. H. Fecteau and D.P. Munoz, "Salience, relevance, and firing: a priority map for target selection," *TREN*, vol. 10, no. 8, pp. 382–390, 2006.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[4] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[5] N. Murray, M. Vanrell, X. Otazu, and C.A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 433–440.

[6] Simone Frintrop, Erich Rome, and Henrik I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.*, vol. 7, no. 1, pp. 6:1–6:39, Jan. 2010.

[7] 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, pp. 1–20, 2008.

[8] Neil D. B. Bruce and John K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[9] Xiaodi Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[10] O. Le Meur, D. Thoreau, P. Le Callet, and D. Barba, "A spatio-temporal model of the selective human visual attention," in *IEEE International Conference on Image Processing*, 2005, vol. 3, pp. 1188–1191.

[11] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2005.

[12] Charles E. Connor, Howard E. Egeth, and Steven Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology*, vol. 14, pp. 850–852, 2004.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2009, pp. 2106–2113.

[14] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 438–445.

[15] A. Borji, D.N. Sihite, and L. Itti, "What/where to look next? modeling top-down visual attention in complex interactive environments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 523–538, 2014.

[16] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[17] SensoMotonic Instruments, "Smi eye tracking glasses," Tech. Rep., http://eyetracking-glasses.com, 2016.

[18] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.

[19] N. Anantrasirichai, K.A.J. Daniel, L. Gilchrist, J. Burn, and D. Bull, "Visual priority maps for biped locomotion," *submitting to PAMI*, 2016.

[20] R.J. O'Callaghan and D.R. Bull, "Combined morphological-spectral unsupervised image segmentation," *IEEE Transactions on Image Processing*, vol. 14, no. 1, pp. 49–62, 2005.

[21] N. Anantrasirichai, Iain D. Gilchrist, and David R. Bull, "Fixation identification for low-sample-rate mobile eye trackers," *submitting to ICIP2016*, 2016.

[22] B. W. Tatler, R. J. Baddeley, and I. D Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision Research*, vol. 45, pp. 643–659, 2005.

[23] A. Borji, H.R. Tavakoli, D.N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *IEEE International Conference on Computer Vision*, 2013, pp. 921–928.

[24] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *IEEE International Conference on Computer Vision*, 2013, pp. 1153–1160.

[25] Erkut Erdem and Aykut Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 1–20, 2013.