Siebert, J. P., Ozimek, P., Balog, L., Hristozova, N. and Aragon-Camarasa, G. (2018) Smart Visual Sensing Using a Software Retina Model. IROS2018 Workshop: Unconventional Sensing and Processing for Robotic Visual Perception, at 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, Madrid, Spain, 05 Oct 2018.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/183359/

Deposited on: 8 April 2019

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Smart Visual Sensing Using a Software Retina Model

Jan P. Siebert[1,2] and Piotr Ozimek[1]and Lorinc Balog[1]and Nina Hristozova[1]and Gerardo Aragon-Camarasa[1]

***Abstract*— We present an approach to efficient visual sensing and perception based on a non-uniformly sampled, biologically inspired, software retina that when combined with a DCNN classifier has enabled megapixel-sized camera input images to be processed *in a single pass*, while maintaining state-of-the recognition performance.**

I. INTRODUCTION

A key issue in designing robotics systems is the cost of an integrated camera sensor that meets the bandwidth/processing requirement for many advanced robotics applications. Lightweight visual sensing is especially important for many applications, such as SLAM in autonomous aerial vehicles, or for wearable camera devices intended for egocentric perception applications. Even in conventional robotics tasks such as, grasping and manipulation, both the sheer visual data rate to be processed in real-time and the need for data efficiency when using Deep Learning technology present significant challenges. As DL networks become ever larger, more sophisticated, and correspondingly more computationally expensive to train, the need for data efficiency is becoming accordingly more critical.

To address the above issues, we have been investigating biologically motivated foveated vision algorithms based on the visual processing architectures found in mammals. This evolutionary development appears to reduce visual load by around two orders of magnitude. Based on this observation, we have developed a foveated visual architecture that implements a functional model of the retina-cortex mapping Fig.1 Right. In our prior research, this retina model served to produce feature vectors that were matched/classified using conventional methods[1]. Our software retina and mapping has been adapted to serve as a data-reducing/normalising preprocessor for classification and interpretation by means of Deep Convolutional Neural Nets (DCNN), [2], [3], allowing megapixel-sized camera input images to be processed *in a single pass*.

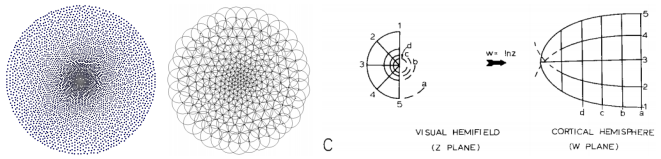

Fig. 1. **Left:** Gaussian receptive fields on top of a retina tessellation, taken from [1]. **Centre:** The 4196 node tessellation used in this paper. **Right:** A schematic view of the retino-coprtical mapping, taken from [4].

*This work was supported by the EU funded *Human Brain Project* and also the Innovate UK funded *iSee* project.

[1]School of Computing Science, University of Glasgow, Scotland, UK. [2]Correspondence: paul.siebert@glasgow.ac.uk

II. APPROACH

The retina model employed in this paper was originally developed by Balasuriya [1] (and later optimised by Ozimek [2]) who investigated developing a visual architecture that integrates feature extraction and gaze control based on a self-organised software retina. To generate the retina tessellation without local discontinuities, distortions or other artefacts, a self-similar neural network is used to define retina sampling locations as described by Clippingdale & Wilson [5], as illustrated in Fig.1 Left. This method relies on a network of N nodes jointly undergoing random translations to produce a tessellation with a near-uniform dense foveal region that seamlessly transitions into a sparse periphery. Each node in the resultant tessellation defines the location of a receptive field's centre. Each receptive field has a Gaussian response profile, the standard deviation of which scales linearly as an (inverse) function of local node density and this in turn scales inversely with eccentricity, visualised in Fig.1 Centre. This scaling balances between introducing aliasing at the sparsely sampled peripheries and super-Nyquist sampling at the densely sampled foveal region [2].

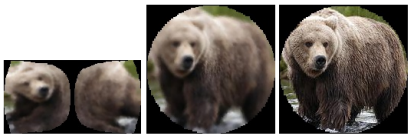

Fig. 2. An example of a transformed *Brown Bear* image: Left Cortex image, Centre: Retina Back-projection, Right: input image.

The values sampled by the receptive fields are then stored in an *imagevector*, comprising a one-dimensional array of intensity values which are input to the remainder of his visual processing chain and are also used to feed the processing pipeline in this work. To be compatible with current DCNN visual processing networks, which require regular image input matrices, we create a *cortical image*, Fig. 2 Left, by projecting the imagevector intensities via Gaussian kernels centred on the *polar-transformed* retina sampling locations.

The cortical image mapping should ideally be *conformal*, i.e. preserve local angles and maintain a fairly uniform receptive field density, while preserving local information captured by the retina without introducing any artefacts. These criteria must be satisfied to enable the convolution kernels of DCNNs to extract features from the resultant cortical image. While Swartz [4] reports modelling the retino-cortical mapping using a modified log-polar transform, we obtained a good mapping experimentally via a modified polar transform that confers a degree of scale and rotation invariance in addition to a substantial degree of visual data reduction. It is also possible to invert the image vector via *back-projection* to

produce an equivalent *retina* image, shown in Fig.2 Centre, c.f. the full-resolution input field-of-view in Fig. 2 Right.

III. RECENT WORK

We reported a first pilot study [2] demonstrating a functional retina-integrated DCNN implementation which confirmed that DCNNs can be trained to recognise objects in cortical space and yield efficiency gains: Using a 4K node retina the method reduced the visual data by ~$\times$7, the input data to the CNN by 40% and the number of CNN training epochs by 36%. In this case the gains came at the expense of a slightly reduced F1 score of 0.80 when applying the full retino-cortical transform in a 4-way classification task, compared to an identical network yielding an F1 score of 0.86 for the same data set, but trained and classified using full-resolution images.

Having demonstrated the basic viability of the retino-cortical mapping-DCNN approach, we implemented a 50K node *high-resolution* retina capable of sampling a 930×930 pixel input image to produce a cortical image of around 150K pixels, yielding a visual data reduction of ~$\times$16.7 and a network input reduction of ~$\times$5.8[3]. A GPU accelerated version of this 50K node retina[6] generates a cortical image in ~13ms when executing on an NVIDIA GTX1080TI GPU, running CUDA C codes.

By pre-processing images collected by means of Tobii Pro 2 eye-tracking glasses [7], to control the fixation locations of a software retina model, we were able to demonstrate [8] that we can reduce the input to a custom designed DCNN architecture (based on DeepFix[9]) by $\times$3 (subsequently improved to $\times$10.8 [10]), reduce the required training time by $\times$3.8 and obtain over 98% classification rate when training and validating the system on a database of over 26,000 images of 9 object classes comprising common supermarket products. In this experiment, our objective was to allow a human operator to collect appropriate training data for a foveated egocentric perception[11], [3], [12] system simply by looking at objects. These objects may then be recognised in images collected by a human observer using eye tracking glasses, or a machine observer equipped with a saliency model to direct visual gaze.

We have also demonstrated the 4K node retina running on an Apple iPhone[13]. This comparatively small retina samples a patch in an image captured by the iPhone's camera and SIFT descriptors are extracted from the cortical image to direct the next retinal fixation location of the next in conjunction with a simple inhibition of return algorithm. The iPhone can therefore serve as an autonomous cortical image capture device for tasks such as object appearance learning.

The GPU accelerated implementation of the 50K node retina has been used to demonstrate real-time gaze control in a target tracking task[14]. In this experiment we trained a retina pre-processed DCNN pipeline with example tracking data based on a centroid colour tracker following an orange coloured target set against a dark background. The DCNN system learned to drive the Baxter Research Robot's wrist camera using pan-tilt signals learned from the cortical images input to a custom DCNN we designed. Accordingly our DCNN was able to regress, directly in real time, from cortical space, the appropriate pan-tilt activation to allow the robot's wrist camera to track the coloured target.

Finally, in order to improve the overall data efficiency of the retina-DCNN combination, we have implemented a gamut of colour and intensity retinal ganglion P-cells[15] based on a model described by [16]. These include Red-Green & Blue-Yellow difference cells, and Red-Green & Blue-Yellow single and double opponent cells. We have also implemented the classical intensity difference of Gaussian response cells. These have been structured to generate the pairs of response outputs corresponding to rectified +ve & -ve responses. As consequence these cells appear to exhibit the first stages of figure-ground separation in their responses to appropriate intensity and colour differential inputs and we anticipate this will simplify learning segmentation and object boundaries.

IV. WORK IN PROGRESS

Our research group is currently tackling the fundamental issue of how best to couple the retina directly to the DCNN to obtain the maximum data efficiency, avoiding the need to construct a cortical image by interpolation. Large-scale validation of the 50K node retina is also under investigation, based on the EPIC Kitchens dataset [17], as are methods for efficiently producing very large retinas able to process images of the order of 25Mpx. This effort is combined with integrating the retina approach with the latest DCNN architectures for robot vision, such as Deep6DPose [18]. Our intention is to process images from a robot on-wrist camera, such as the provided by the Baxter Research Robot, or an experimental in-palm camera mounted on the Smart Grasping System, developed by The Shadow Robot Company. We are also porting the retina to the Android smartphone platform and produce self-contained smartphone-based retina sensors for robotics and egocentric perception applications and a convenient method for capturing training data by non-expert users. We are constructing interfaces to commercial high-resolution pan-tilt security cameras to allow these systems to serve within active vision systems using retina processing and cortex-based gaze control algorithms. To manage the data generated by these retina-supported camera systems, we are developing tools to allow editing and formatting prior to training DCNN systems.

V. CONCLUSIONS

We have demonstrated that it is possible to make substantial data efficiency gains in terms of training computation, network sizes and inference rates by pre-processing images using a biologically inspired retina-cortex mapping that affords both visual data reduction and also a degree of scale and rotation invariance. Our ongoing investigations aim to realise fully the potential gains for this approach, its integration within mainstream robot vision DCNNs and underpin practical low-cost visual sensors for autonomous systems.

REFERENCES

[1] Balasuriya, S.: A Computational Model of Space-Variant Vision Based on a Self-Organized Artifical Retina Tesselation. PhD thesis, Department of Computing Science, University of Glasgow, Scotland (March 2006)

[2] Ozimek, P., Siebert, J.: Integrating a Non-Uniformly Sampled Software Retina with a Deep CNN Model. In: BMVC 2017 Workshop on Deep Learning On Irregular Domains. (September 2017)

[3] Ozimek, P., Balog, L., Wong, R., Esparon, T., Siebert, J.P.: Egocentric Perception using a Biologically Inspired Software Retina Integrated with a Deep CNN. In: ICCV 2017 Workshop on Egocentric Perception, Interaction and Computing. (October 2017)

[4] Schwartz, E.L.: Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. Biological Cybernetics **25**(4) (1977) 181–194

[5] Clippingdale, S., Wilson, R.: Self-similar neural networks based on a kohonen learning rule. Neural Networks **9**(5) (1996) 747–763

[6] Balog, L.: A GPU accelerated software retina. Master's thesis, School of Computing Science, University of Glasgow, Glasgow, Scotland UK (2017)

[7] AB, T.: Tobii pro glasses 2: Users manual (Nov 2017)

[8] Hristozova, N.: Dissertation using eye-tracking glasses for training dcnns (March 2018)

[9] Kruthiventi, S.S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing **26**(9) (Sept 2017) 4446–4456

[10] Shaikh, H.A.: A Data Efficient Retino-Cortical Image Transformation Mapping . Master's thesis, School of Computing Science, University of Glasgow, Glasgow, Scotland UK (2018)

[11] Siebert, J., Schmidt, A., Aragon-Camarasa, G., Hockings, N., Wang, X., Cockshott, W.P.: A Software Retina for Egocentric & Robotic Vision Applications on Mobile Platforms. In: ECCV 2016 Workshop on Egocentric Perception, Interaction and Computing. (October 2016)

[12] Hristozova, N., Ozimek, P., Siebert, J.P.: Effcient Egocentric Visual Perception Combining Eye-tracking, a Software Retina and Deep Learning. In: ECCV 2018 Workshop on Egocentric Perception, Interaction and Computing. (September 2018)

[13] Wong, R.: A Smartphone Software Retina. Master's thesis, School of Computing Science, University of Glasgow, Glasgow, Scotland UK (2017)

[14] Boyd, L.: A retina-based vision system for motion control of the baxter robot (March 2018)

[15] Esparon, T.: An investigation of a software retina for robot vision. Master's thesis, School of Computing Science, University of Glasgow, Glasgow, Scotland UK (2017)

[16] Gao, S., Yang, K., Li, C., Li, Y.: Color constancy using double-opponency. IEEE Transactions on Pattern Analysis & Machine Intelligence **37**(10) (Oct. 2015) 1973–1985

[17] Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV). (2018)

[18] Do, T., Cai, M., Pham, T., Reid, I.D.: Deep-6dpose: Recovering 6d object pose from a single RGB image. CoRR **abs/1802.10367** (2018)