Object recognition in egocentric videos with saliency-based non uniform sampling and variable resolution space for features selection

Vincent Buso, Jenny Benois-Pineau, Iván González-Díaz

Object recognition in egocentric videos with saliency-based non uniform sampling and variable resolution space for features selection

Vincent Buso
University of Bordeaux
LaBRI, UMR 5800, Bordeaux, France
vbuso@labri.fr

Iván González-Díaz
Universidad Carlos III de Madrid
Leganés, 28911, Madrid, Spain
igonzalez@tsc.uc3m.es

Jenny Benois-Pineau
University of Bordeaux
LaBRI, UMR 5800, Bordeaux, France
benois-p@labri.fr

January 8, 2015

1 Introduction

Since recently a new video content is massively coming into practice: the egocentric videos recorded by body-worn cameras. In the context of this work which is the behavioral study patients with Alzheimer disease, this kind of video content allows for a close-up view of instrumental activities of daily living (IADL). In parallel, automatic extraction of visually salient areas from this kind of video content is a strong research direction since it brings the focus of attention to objects interacted (manipulated, observed) during IADLs. Recognition of manipulated objects is a key cue for an automatic activity assessment.

In this work we describe our approach for object recognition using visual saliency modeling. As presented in Figure 1, we build our model on the well-known BoW paradigm [3], and propose a new approach to add saliency maps in order to improve the spatial precision of the baseline approach. For saliency maps computations, we use the model described in [1]. Finally we use a non-linear classifier to detect the presence of a category in the image.

In this research, the contribution of saliency is twofold.

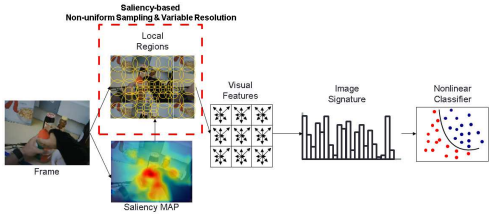

Figure 1: Processing pipeline for the saliency-based object recognition in first-person camera videos

First, it controls how and where circular local patches are sampled in an image for descriptor computation. Second, it controls the spatial resolution at which the features are computed. Our aim is to emulate the retina in the Human Visual System (HVS) where cells in charge of foveal and peripheral vision work at different spatial resolutions [8]. In the following section, we briefly describe our approach, then are presented the experiments and results. The last section provides conclusions and perspectives.

2 Saliency-based object recognition

In this section we describe how we emulate retina properties at different spatial resolutions [8] using Non-Uniform Sampling of features and their selection at Variable Spatial Resolution (NUS+VSR).

Due to varying morphology of its neurons, the human eye retina simultaneously allows for a high-resolution and detailed perception at the visual field associated to the fovea, and a low-resolution one in the peripheral visual field [8]. This leads to a non-uniform spatial resolution image analysis, commonly known as *foveation* [2].

In order to model the difference between visual fovea and peripheral pathways found in human retina [8], our approach combines space-variant sampling [7] and multi-resolution image foveation [5]. Fig. 2 illustrates the approach.

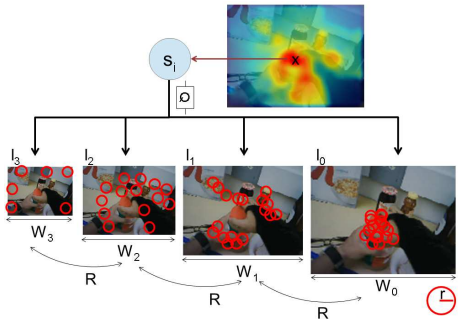

Figure 2: Representation of our Visual Fields with Variable Spatial Resolution and Non-uniform sampling

2.1 Variable Spatial Resolution (VSR)

In order to provide a multi-resolution analysis of an input image, we first discretize the resolution space by generating a multi-scale Gaussian image pyramid of L levels. Lower levels are meant to represent foveal vision whereas upper ones model peripheral vision. As shown in Fig. 2, we define the *resolution factor* ρ that stands for the ratio between the widths of two contiguous images in the pyramid $W_l = \frac{W_{l-1}}{\rho}$, where l denotes the level in the pyramid.

Then, using values in the saliency map, we can compute the saliency value of each local circular patch n. Depending on this value $s_n \in [0, 1]$ we assign each local patch to a particular level of the pyramid. In our case, this is done by simple linear Quantization ($Q(s)$ in Fig. 2). Intuitively, high-saliency patches associated with foveal vision will follow high-resolution visual processing pathways, whereas low-saliency patches associated to the peripheral vision will utilize the low-resolution (higher) levels of the pyramid.

Our approach discretizes hence the resolution space, what differs from previous works towards foveated video displays [2, 5].In order to establish an analogy with the HVS, we are modeling the foveal visual field as a high-resolution pathway that pays attention to small image details, whereas peripheral vision path acquires information at a lower resolution, thus focusing on coarse visual patterns.

2.2 Non-uniform Sampling (NUS)

Here we introduce the pruning process that filters out visual information in order to provide more compact image representations focusing on areas with high saliency.

We follow a similar approach to [7], in which a Weibull cumulative distribution was proposed to perform random sampling based on saliency values. A Weibull distribution has two tunable parameters, namely: a shape parameter k, and a scale parameter λ. The former manages the influence of the saliency over the sampling process, for a given scale, the latter controls the number of sampled points. Hence, since we aim at improving classification results while avoiding any additional processing burden, we have cross-validated the value of k, and developed a lower bound on the number of sampled points that allows to analytically derive the corresponding value of λ that keeps the computational complexity constant.

3 Experiments and results

In this section we describe the dataset and assess our proposed perceptual models in an object recognition problem.

We have used the publicly available *ADL* dataset [6] where only the object labeled as 'active' were considered. The recordings of IADLs are made at different home locations which corresponds to an unconstrained, thus very challenging scenario.

Regarding the experiments, we have compared several approaches to our NUS+VRS model. We have included two reference methods for comparison: a) *Baseline BoW*

Table 1: A comparison of various configurations of Saliency-based Object Recognition for the whole (44) and reduced (10) sets of categories in ADL dataset.

Algorithm/mAP (%)	ADL (44 cat)	ADL (10 cat)
BoW	10.36	30.75
BoW + GT Masks	13.85	43.79
NUS [7]	10.81	31.51
NUS+VSR	11.74	35.57

with a vocabulary size of 4000 visual words and b) *BoW + GT Masks* in order to evaluate the theoretical limit of the use of saliency maps. Regarding our NUS+VSR module described in Section 2 we evaluated: a) *NUS*which consists of BoW with the NUS module described in [7], but we have introduced the upper bound on the complexity and b) *NUS+VSR* where we add the VSR module (section 2.1) to the previous approach.

Results for every method are contained in Table 1, which allows us to draw very interesting conclusions. Although the NUS already enhances the basic BoW, the VSR approach still provides notable improvements to the system performance. This is a nice consequence of the spatial resolution adaptation to the foveal and peripheral vision, and raises the need of independent and different scale processing paths for the objects of interest and the context in an image.

4 Discussion

In this work we have presented a way to emulate properties of the human visual system for the challenging task of object recognition in egocentric videos. In particular we showed how saliency can be used for the modelling of visual fields thanks to a Variable-Resolution Space and Non-Uniform Sampling. Regarding the object-recognition task in egocentric videos, we have shown that our biologically inspired model helps improving performances. In the future we plan to continue exploring the benefits of saliency throughout recognition tasks.

References

[1] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet. A metric for no-reference video quality assessment for hd tv delivery based on saliency maps. In *IEEE International Conference on Multimedia and Expo*, july 2011. 1

[2] E.-C. Chang, S. Mallat, and C. Yap. Wavelet foveation. *Applied and Computational Harmonic Analysis*, 9(3):312 – 335, 2000. 2

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004. 1

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, Nov. 1998.

[5] J. S. Perry and W. S. Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *In Human Vision and Electronic Imaging, SPIE Proceedings*, pages 57–69, 2002. 2

[6] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 2

[7] E. Vig, M. Dorr, and D. Cox. *Space-Variant Descriptor Sampling for Action Recognition Based on Saliency and Eye Movements*, pages 84–97. Springer, Firenze, Italy, 2012. 2, 3

[8] B. A. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., 1995. 1, 2