EgoMemNet: Visual Memorability Adaptation to Egocentric Images

Marc Carné and Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Terrassa, Catalonia/Spain
marc.carne.herrera@estudiant.upc.edu
xavier.giro@upc.edu

Petia Radeva
Universitat de Barcelona
Barcelona, Catalonia/Spain
petia.ivanova@ub.edu

Cathal Gurrin
Insight Centre for Data Analytics
Dublin, Ireland
cathal.gurrin@dcu.ie

Abstract

This work explores the adaptation of visual memorability prediction for photos intentionally captured by handheld cameras, to images passively captured from an egocentric point of view by wearable cameras. The estimation of a visual memorability score for an egocentric images is a valuable cue when filtering among the large amount of photos generated by wearable cameras. For this purpose, a new annotation tool and annotated dataset are presented, used to fine-tune a pre-trained convolutional neural network.

1. Introduction

The popularization of wearable cameras has increased the interest of an automatic filtering of the large amounts of images and videos they generate. These devices are typically used throughout the day for lifelogging purposes, that is, to generate a digital memory of their users. Capturing passively during the whole day, the camera generates up to 2000 images per day. A question arises what would be the memorable images from this large set of data. Therefore, this work explores the automatic prediction of the memorability (defined by multiple factors) of the captured data. We extend a previous work on visual memorability [5] to the specific domain of egocentric vision.

Egocentric images present a particular composition that is different from photos taken with a handheld camera. For this reason, we present in this paper a new dataset of egocentric images with an annotated memorability score.

Filtering large sets of images generated by wearable cameras has received attention of several researchers in the literature. A first approach is clustering images based on low-level perceptual features, and choosing a single representative image for each cluster for summarization [2]. Another option is to select only those images, whose quality exceeds a minimum threshold [3]. Other authors have moved into even more abstract concepts such as intentionality [7] or novelty [1].

Visual memorability has been largely explored in the past by psychologists, and more recently by computer vision scientists. Contributions can be grouped into those considering images at a global scale [4], or those focusing on a more local scale [6, 5]. State of the art performance is achieved by *MemNet* [5], a convolutional neural network fine-tunned with 60,000 images annotated with the visual memory game introduced in [4].

2. Egocentric Image Annotation

The proposed model for memorability prediction of egocentric images required the annotation of a new dataset to fine-tune the existing *MemNet* model.

Our annotation tool is a new implementation inspired by the visual memory game proposed in [4]. During the game, the user is shown a sequence of images with a blank frame (small black square at the center of a white background) between them, and is asked to press a key whenever a repetition is detected. The image sequence contains two type of images: *target* images which are to be annotated, and *fillers* which are to be ignored when estimating the final annotation score. Some fillers have a vigilance role in order to control the user attention and quality of his/her work. Targets appear only twice (at random distance between 8 and 40 frames apart) in the sequence.

The detection rate of each target image through all annotators defines its *memorability score*. This value estimates the probability of one person remembering having seen an image when this image is shown for a second time after a short period since the first view.

The main challenge when building a dataset for egocentric visual memorability is the large amount of similar images within a temporal neighborhood, which may cause false detection. We try to reduce the chances of these erroneous detection by using two different sets of images, one for target images and a second one for fillers. Figure 1 provides some samples of both targets and fillers.

The set of targets contains 50 images (captured with Autographer) used as targets for the annotation tool. This col-

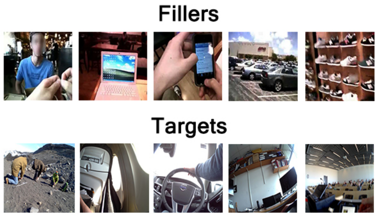

Figure 1. Examples of the images used in the game: fillers (from UTEgocentric dataset) on top and targets (from Anonymous dataset) on bottom.

lection was built from a uniform sampling from a lifelogging record of 25 days, which corresponds to 16,000 images.

The set of fillers was built by sampling frames from the UTEgocentric dataset [7], which contains 4 videos, each about 3-5 hours, from head-mounted cameras. Videos were sampled, extracting frames at a fixed interval of time, in this case every 30 seconds, to simulate the capture by a low sampling rate camera and were used as fillers for the annotation tool.

3. Results

The annotation tool and datasets presented in the previous sections were used to annotate a total of 50 egocentric images by 25 different users. These data were used to finetune *MemNet* convolutional neural network. The dataset was split into 40 images for train and validation, and the remaining 10 images for test.

Given the reduced amount of images available for training, we explored three different strategies for data augmentation: no augmentation, Spatial Data Augmentation (SDA) and Temporal Data Augmentation (TDA).

To evaluate the performance of each model, we used MSE (Mean Square Error) and Spearman's Rank Correlation, comparing relative positions of a set of items between two list, both between the ground truth and predicted scores.

As a result, seven models were created and compared with *MemNet* [5]: the first model (*fc7*) only fine-tuned the last layer before the regressor, *SDA-fc7* corresponds to the spatial augmentation case, and *TDA-fc7*, *TDA-fc6*, *TDA-conv5*, *TDA-conv5+fc6+f7* and *TDA-fc6+fc7* corresponds to cases of temporal augmentation up to the specified fully connected (fc) or convolutional (conv) layers. Result on the test images are presented in Table 1.

Model	Rank correlation	MSE
MemNet [5]	0.7727	0.00846
fc7	0.7606	0.01574
SDA-fc7	0.7182	0.01977
TDA-fc7	0.8152	0.01635
TDA-fc6	**0.8394**	0.01085
TDA-conv5	0.8091	0.00578
TDA-conv5+fc6+fc7	0.7909	0.00737
TDA-fc6+fc7	0.7485	**0.00486**

Table 1. Quantitative evaluation.

The results depicted in Table 1 indicate that those models fine-tuned with egocentric images outperformed *MemNet* when temporal data augmentation is considered. According to our experiments, fine-tunning only layer fc6 is the best option for this set up. This best configuration is released together with this publication as *EgoMemNet*, simultaneously with the annotation tool and the dataset to facilitate the replication and extension of the presented results.

4. Conclusions

This work has presented a successful adaptation for visual memorability prediction, from the handheld camera domain to the wearable camera one, taking *MemNet* a as baseline. The performance of an existing model for visual memorability has been clearly improved with a small set of 50 annotated images augmented in the temporal domain and fine-tunning the first of the fully connected layers.

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3297–3304. IEEE, 2011.

[2] M. Bolanos, R. Mestre, E. Talavera, X. Giro-i Nieto, and P. Radeva. Visual summary of egocentric photostreams by representative keyframes. In *Multimedia & Expo Workshops (ICMEW), IEEE International Conference on*, pages 1–6, 2015.

[3] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 259–268. ACM, 2008.

[4] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1469–1482, 2014.

[5] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2390–2398, 2015.

[6] A. Khosla, J. Xiao, A. Torralba, and A. Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems*, pages 305–313, 2012.

[7] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *Computer Vision–ECCV 2014*, pages 282–298. Springer, 2014.