Dublin's Participation in the Predicting Media Memorability Task at MediaEval 2018

Alan F. Smeaton [1], Owen Corrigan [1], Paul Dockree [2], Cathal Gurrin [1], Graham Healy [1],
Feiyan Hu [1], Kevin McGuinness [1], Eva Mohedano [1], Tomás Ward [1]
[1] Insight Centre for Data Analytics, Dublin City University,
[2] School of Psychology and Trinity College Institute of Neuroscience, Trinity College Dublin, Ireland
alan.smeaton@dcu.ie

ABSTRACT

This paper outlines 6 approaches taken to computing video memorability, for the MediaEval Predicting Media Memorability Task. The approaches are based on video features, an end-to-end approach, saliency, aesthetics, neural feedback, and an ensemble of all approaches.

1 INTRODUCTION

In our work we seek to explore theories from psychology and neuroaesthetics, which may guide predictors for memorability of visual media. Two caveats are that most of the ideas from neuroaesthetics come from perception of visual art or artificial experimental stimuli, rather than real life scenes so these ideas might not translate. The second caveat is that over and above the aesthetics of the video or keyframes, we cannot control for the semantic content or the emotional salience of the imagery for the viewer just as we cannot control for the viewer's attention or concentration while initially viewing or subsequently trying to remember the video.

Our first principle is the idea that aesthetically pleasing features are driven by Gestalt principles [10] including grouping, symmetry and lines of good continuation. In each case, items in a scene are bound together into coherent groups or continuous unbroken forms by our visual system. According to Ramachandran [7], these Gestalt principles are driven by neural mechanisms in our perceptual system that trigger the brain's reward system so that our attention is reflexively drawn to these features. There is also some evidence that grouping of visual features not only increases attention but also benefits visual working memory [6].

Our second principle, and in opposition to processing a coherent whole, is that images that show distinctive figure/ground arrangements may also capture attention thus promoting memorability. So, another of Ramachandran's laws of neuroaesthetics is "isolation" in which a key visual feature has exaggerated importance and stands out from the surrounding information [8].

Although these aesthetic features are intrinsic qualities in images that capture attention, it is less clear how they affect memorability. However superior attention based on these qualities should increase encoding of the videos and hence improve memorability. Thus a key prediction based on these principles is that a U-shaped relationship should emerge in which the most globally coherent video images and the most locally distinctive images should both be more memorable compared to the video frames that fall in-between these extremes – i.e. those that are neither particularly globally coherent nor locally distinctive.

This work in this paper was carried out in the context of the 2018 MediaEval Predicting Media Memorability task and we refer the reader to the task description for prior art [1].

2 RUNS SUBMITTED

2.1 Machine Learning with Pre Computed Features

In this run, we evaluated the performance of a neural network to run on the precomputed features provided by the task organisers. These features include C3D features, HMP, HOG Descriptors and more. The complete list can be found in [1]. To merge these different features, we simply flattened them into one long vector. Using this as an input, we trained a Multi Layer Perceptron which would output a probability. We tested a number of architectures and found in testing that using 3 layers was optimal.

2.2 An End-to-end System

For our end-to-end system we used 3 keyframe images from the raw videos as inputs. At each epoch, we selected one frame randomly from the video as a form of data augmentation. For the architecture, we tried two standard models: VGG16 [9] and Resnet18 [2]. We modified these networks by changing the output to target a single variable, memorability, instead of matrix of class probabilities. We also investigated using different numbers of dense layers after the convolutional layers. Surprisingly, we found that using a single layer with VGG16 gave the best results. Our loss function was mean squared error, and we used a gradient descent optimizer.

2.3 Using Video and Image Saliency

Visual saliency models generate a probability map highlighting image regions that most attract human attention. Here, this information is explored for the task of predicting media memorability. More precisely, a saliency map for each frame of video is computed with the SalGAN model [5].

The maps are used to spatially weight the activations of the last convolutional layer of Inception-v3 pre-trained on Imagenet. For that, video frames are resized to 300×300 resolution, and forwarded to Inception-v3 to generate convolutional volumes of $7 \times 7 \times 2048$ (the first two dimensions correspond to the spatial resolution, and the last one the number of channels or depth of the layer).

Saliency maps are downsized to 7×7, normalised to contain values between 0-1, and element-wise multiplied to the convolutional activations. Global average pooling is applied on the channel

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

dimension to obtain a final representation of 2048 dimensions. The hypothesis here is that the denser the saliency map the more human attention the images draw, and consequently the more memorable they may be.

This 2048 long vector was then fed into a neural network, similar to how precomputed features were used in Section 2.1.

2.4 Using Neural Approach

In this approach we used human reaction to a second viewing of a video keyframe, to train a classifier for memorability, a true human-in-the-loop experiment. The middle frame was extracted for each video clip in the test set and a participant was shown these images at high speed (4 Hz) on a computer screen while simultaneously recording their EEG (Electroencephalography) signals.

Each of the 2000 test set extracted images were presented twice. Following completion of the first viewing, EEG signals were band-passed between 0.5 Hz and 10 Hz, re-referenced to a common average reference and the mean voltage between 300ms and 600ms following each image presentation calculated for the Pz channel (baselined to -250ms to 0 ms prior to image presentation). The participant then viewed the images a second time with similar EEG data recording and processing and the values averaged for the two presentations of each image, which formed the submission scores.

These parameters were selected as they are known to correspond both to a time region and electrode location in which a P300 event-related potential in this type of task is typically observed where attention is elicited [3]. The rationale is that high amplitude P300 responses correspond to imagery which is visually attentive and thus potentially more memorable which should also stimulate visual working memory [6]. We then computed the pearson correlation between the P300 signals and the memorability scores to evalute the performance of this feature.

2.5 Computing Visual Aesthetics

A final technique we incorporated was to use our own version of an image aesthetics classifier as described in [4], instead of the values provided by the task organisers. This maps back to our guiding principles driven by neuroaesthetics, described earlier.

2.6 An Ensemble of All Techniques

In each of the approaches above we made predictions for the entire training set, as well as the entire testing set after training had completed. One limitation to note is that due to the time consuming nature of EEG labelling in Section 2.4, only a subset of the training dataset (2,000 videos) was used in this ensemble run. We used predictions from each of the above approaches, and trained a linear model on this subset of the training data to identify which were the most important predictors. We then used these weights to combine the values on the test set, which generated this run.

3 RESULTS, CONCLUSIONS AND FUTURE PLANS

The performance results of our submissions are shown in Table 1 and illustrated in Figure 1.

Run type	Ensemble	Features	End-to-end	Saliency	Neural
Short Term Memorability					
Spearman	-0.018	0.051	0.055	-0.015	-0.027
Pearson	-0.019	0.026	0.085	-0.015	-0.031
MSE	0.0089	0.0069	0.0069	0.0073	0.0089
Long Term Memorability					
Spearman	0.039	0.037	0.017	0.007	-0.024
Pearson	0.021	0.016	0.032	0.006	-0.024
MSE	0.0207	0.0205	0.0207	0.0208	0.0207

Table 1: Results

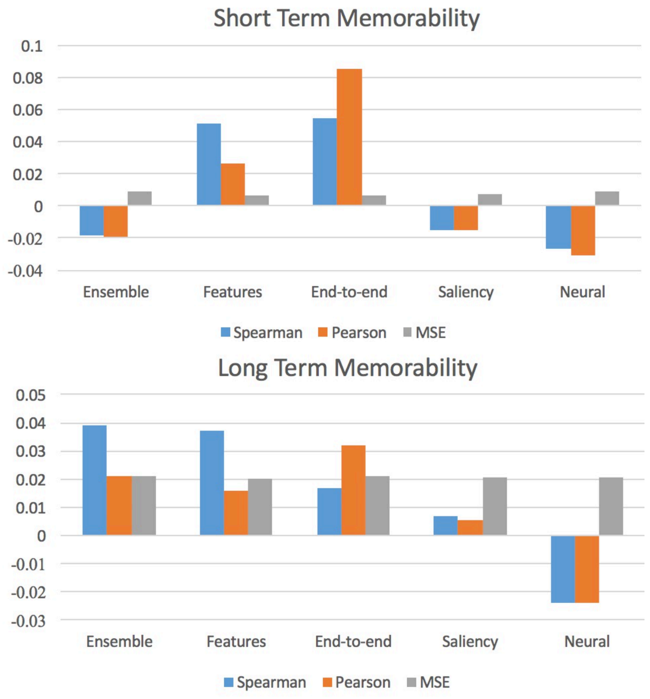

Figure 1: Performance for memorability classification

The results show that the run based on direct neural/EEG feedback from the human participant was the worst, as expected, and part of the reason might be because training was done with on only 2,000 images, with only one participant. It is definitely worth scaling up this approach to see performance with more data.

The run based on our saliency was a bit better than the neural run, especially for long-term memorability. The ordering of runs by performance among the provided features, ensemble and end-to-end submissions has contradictions across runs, across long vs. short term memorability, and across the metric used but the end-to-end seems to have performed best, which is surprising.

Overall, our results seem poor for the above reason or because of insufficient tuning of parameter settings in our experiments.

ACKNOWLEDGMENTS

This work was partially supported by Science Foundation Ireland under the SFI Research Centres Programme grant number SFI/12/RC/2289.

REFERENCES

[1] R. Cohendet, C.-H. Demarty, N.Q. Duong, M. Sjöberg, B. Ionescu, and T.-T. Do. 2018. MediaEval 2018: Predicting Media Memorability. In *Proc. of the MediaEval 2018 Workshop, Sophia-Antipolis, France*. CEUR-WS, Sophia-Antipolis, France, 29–31.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Las Vegas, United States, 770–778.

[3] Graham Healy, Tomas Ward, Cathal Gurrin, and Alan F. Smeaton. 2017. Overview of NTCIR-13 NAILS Task. In *Proceedings of the NTCIR-13 NAILS (Neurally Augmented Image Labelling Strategies)*. National Institute of Informatics, Japan, Tokyo, Japan, 380–383.

[4] Feiyan Hu and Alan F. Smeaton. 2018. Image Aesthetics and Content in Selecting Memorable Keyframes from Lifelogs. In *MultiMedia Modeling - 24th International Conference, MMM, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I*. Springer, Bangkok, Thailand, 608–619.

[5] Junting Pan, Cristian Canton-Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i-Nieto. 2017. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. *CoRR* abs/1701.01081 (2017), 1–9. arXiv:1701.01081 http://arxiv.org/abs/1701.01081

[6] Dwight J. Peterson and Marian E. Berryhill. 2013. The Gestalt principle of similarity benefits visual working memory. *Psychonomic Bulletin & Review* 20, 6 (Dec 2013), 1282–1289.

[7] Vilayanur S Ramachandran. 2012. *The tell-tale brain: A neuroscientist's quest for what makes us human*. WW Norton & Company, 500 Fifth Avenue, New York, New York.

[8] Vilayanur S Ramachandran and Diane Rogers-Ramachandran. 2010. Reading between the Lines. *Scientific American Mind* 21, 4 (2010), 18–20.

[9] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014), 1–14. arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[10] D. Todorovic. 2008. Gestalt principles. *Scholarpedia* 3, 12 (2008), 5345. revision #91314.