LAPI at MediaEval 2017 - Predicting Media Interestingness

Mihai Gabriel Constantin[1], Bogdan Boteanu[1], Bogdan Ionescu[1]
[1]LAPI, University "Politehnica" Bucharest, Romania
{mgconstantin, bboteanu, bionescu}@imag.pub.ro

ABSTRACT

In the following paper we will present our contribution, approach and results for the MediaEval 2017 Predicting Media Interestingness task. We studied several visual descriptors and created several early and late fusion approaches in our machine learning system, optimized for best results for this benchmarking competition.

1 INTRODUCTION

Multimedia interestingness has been studied more and more extensively in recent years, from several perspectives including psychology and computer vision. From a psychological perspective user studies described a correlation between human interest and several other concepts including, but not limited to aesthetics, enjoyment, complexity, novelty [1, 8], while computer vision approaches studied various sets of features and machine learning techniques that are able to predict the interestingness of multimedia shots, based on low-level attributes such as color histograms, SIFT, edge distributions [8] or high-level attribute like composition rules or the presence of certain objects [7].

The MediaEval 2017 Predicting Media Interestingness task [6] creates a benchmarking competition where participants are tasked with the creation of a system that can predict the interestingness of images and video segments annotated by a team of viewers, according to a Video on Demand scenario, where a set of most interesting frames or video shots has to be presented to a certain user. This paper will thus describe our approach for this task.

2 APPROACH

The approach presented in this paper is a continuation of our work described in [3], with the addition of a video interestingness prediction system. The first step in our machine learning system is the extraction of the content descriptors, followed by the learning stage for these content descriptors and their early and late fusion combinations executed on the annotated development dataset. In the final stage we evaluate the best performing combinations on the unlabeled testing dataset. The features used here are presented, along with a detailed description in [3] and are based on the works of [5, 9–11]. These features have been used in several domains connected with interestingness such as aesthetics, photographic compositional rules, color theory etc. For the machine learning algorithm we used Support Vector Machine (SVM) [4] with different parameters and kernels.

2.1 Features

The features used in this system are as follows: Hue, Saturation, Value computed from HSV space (denoted *HSV*), Hue, Saturation, Lightness extracted from HSL space (*HSL*), Colorfulness [5, 9], Hue descriptors (*HueDesc*) [9, 11], Hue models (*HueModel*) [11], Brightness [10, 11], Edge [9–11], Texture [9], RGB entropy (*RGBEntropy*) [9], HSV wavelet (*HSVwavelet*) and average value for the HSV wavelet (*aHSVwavelet*) [5], average HSV values based on the Rule of Thirds (*aHSVRot*) [5], average HSL values for the focus region (*aHSLFocus*) [11], size analysis for the largest five segments (*LargSegm*) [5], centroid placement (*Centroids*) [5], Hue, Saturation, Value and Brightness for the largest segments (*HueSegm, SatSegm, ValSegm, BrightSegm*) [5, 11], color model for the largest segments (*ColorSegm*) [5], coordinates of the segments (*CoordSegm*) [11], mass variance, skewness and contrast between the segments (*MassVarSegm, SkewSegm, ContrastSegm*) [11] and finally a depth of field indicator (*DoF*) calculated according to the method presented in [5].

While for the image subtask each image generated a set of the presented descriptors, for the video subtask we generated two sets of descriptors for each of the individual segments. These two sets of descriptors were generated by extracting the feature set for each frame and then calculating the average value and median value over all the frames in a video segment.

2.2 Data fusion

In both subtasks we used early and late fusion techniques for maximizing out final results. Early fusion combinations consisted of concatenating several features and using the newly created feature as an input for a new training algorithm, while for the late fusion approach we used the confidence output values of several runs and combined them in several strategies, thus generating new confidence outputs.

For the late fusion trials we used 4 strategies: CombMax and CombMin, where we took the maximum and minimum confidence value for each media sample and used them as new outputs, CombSum, where we added up the individual confidence values of the runs and CombMean where the added confidence values were also multiplied with weights distributed according to the rank of the initial system. This weight was calculated as $w = 1/(2^r)$, where the rank r had the value 0 for the best component output classifier, 1 for the second and so on.

2.3 Learning system

The learning system we used was SVM, implemented with the LibSVM library [2], with linear, polynomial and RBF kernels. For the degree, gamma and cost coefficients we used the combinations of values 2^k, where $k \in [-6, ..., 6]$.

3 EXPERIMENTAL RESULTS

As presented in the task overview paper [6], the development dataset consisted of 7396 frames for the image subtask and 7396

MediaEval'17, 13-15 September 2017, Dublin, Ireland

Table 1: Best results on devset for the image and video subtasks and their final result on testset (best testset results are marked in bold)

Subtask	Run	Approach	MAP@10 devset	MAP testset	MAP@10 testset
image	run1	CombMax (HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.0821	0.1791	0.0463
image	run2	CombMax (HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus)	0.0803	0.1789	0.0442
image	run3	CombMean (aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm)	0.0793	**0.1873**	**0.0555**
image	run4	CombMean (HSVWavelet + aHSVWavelet + aHSLFocus and HSV + HSL + aHSLFocus and HSV + MassVarSegm)	0.0793	0.1851	0.0529
video	run1	CombMax (LargSegmMED + ValSegmMED and TextureMED + MassVarSegmMED)	0.0753	0.1937	0.0619
video	run2	CombMax (LargSegmMED + ValSegmMED and TextureMED + MassVarSegmMED and EdgeAVG + TextureAVG)	0.0737	0.1819	0.0564
video	run3	CombMax (EdgeAVG + TextureAVG and HSVAVG + MassVarSegmAVG)	0.0732	0.1937	0.0619
video	run4	CombMean(LargSegmMED + ValSegmMED and TextureMED + MassVarSegmMED and EdgeAVG + TextureAVG)	0.0725	**0.2028**	**0.0732**
video	run5	CombMax (EdgeAVG + TextureAVG and HSVAVG + MassVarSegmAVG and HSLAVG + ColorfulnessAVG)	0.0723	0.1843	0.0571

video segments for the video subtask, while the test dataset had 2435 frames for the image subtask and 2435 video segments for the video subtask. The official metric was mean average precision at 10 (MAP@10), and the organisers also calculated the mean average precision (MAP) for each submitted run. A large number of experiments with different early and late fusion strategies and with different SVM systems were carried out and the best performing combinations were in the last phase run on the testset.

3.1 Experiments on the devset

Our SVM training system used a 10-fold cross-validation approach for choosing the best SVM-feature set combination. Generally, taking into account the MAP@10 metric, the best performing SVM kernel was the RBF kernel. Also another general observation is that the late fusion approaches, especially CombMax and CombMean, outperformed the early fusion combination, while early fusion outperformed learning systems with single descriptors. On the other hand, CombMin and CombSum strategies performed worse than their components with many combinations. Regarding the two descriptor sets for the video subtask (average and median), the results were mixed, some early fusion or single descriptors performing better with the median approach while others performed better when we used the average calculation.

The interestingness confidence score for each shot used for the MAP@10 calculation were extracted as the margin to the decision hyperplane.

Table 1 shows the best results registered on both the image and the video subtasks, and as mentioned earlier the best results were achieved for the late fusion approaches. For the video subtask we used the notation AVG for features that were obtained using average and MED for features that were obtained using median. All the components in Table 1 were trained using the best performing SVM RBF kernel.

For the image subtask the best result on the devset was obtained with a CombMax strategy combining the early fusion outputs of HSV + HSL + aHSLFocus and aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm, with a MAP@10 score on the devset of 0.0821. For the video subtask the best result was a CombMax strategy containing LargSegmMED + ValSegmMED and TextureMED + MassVarSegmMED early fusion outputs, with a MAP@10 score of 0.0753.

3.2 Official results on testset

For the final submission we trained the systems on the entire devset, using the optimal parameters that we found in the previous experiments and tested the resulting systems on the testset.

Table 1 also presents the official results on the testset runs for the combinations we submitted, as returned by the task organisers, with the MAP and MAP@10 scores for each of the runs. For the image subtask we have a best MAP@10 score of 0.0555, obtained by using a CombMean strategy with the outputs of aHSVRot + aHSLFocus and HSV + MassVarSegm + LargSegm. The same system also had the best MAP score - 0.1873. For the video subtask again it was a single system that got both the best MAP@10 and the best MAP score - a CombMean strategy usign the early fusion outputs of LargSegmMED + ValSegmMED and TextureMED + MassVarSegmMED and EdgeAVG + TextureAVG, with a MAP@10 value of 0.0732 and a MAP value of 0.2028.

4 CONCLUSIONS

In this paper we presented several systems that predict media interestingness using content descriptors and early and late fusion approaches. We tested these systems on the MediaEval 2017 Predicting Media Interestingness task and our best results were MAP@10 0.5555 for the image subtask and 0.0732 for the video subtask.

ACKNOWLEDGMENTS

Part of this work was funded by UEFISCDI under research grant PNIII-P2-2.1-PED-2016-1065, agreement 30PED/2017, project SPOTTER

REFERENCES

[1] Daniel E Berlyne. 1960. Conflict, arousal, and curiosity. (1960).

[2] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 27.

[3] Mihai Gabriel Constantin and Bogdan Ionescu. 2017. Content Description for Predicting Image Interestingness. In *International Symposium on Signals, Circuits and Systems - ISSCS 2017*.

[4] Corinna Cortes and Vladimir Vapnik. 1995. Support vector machine. *Machine learning* 20, 3 (1995), 273–297.

[5] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*. Springer, 288–301.

[6] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc QK Duong. 2017. Mediaeval 2017 predicting media interestingness task. In *MediaEval 2017 Multimedia Benchmark Workshop Working Notes Proceedings of the MediaEval 2017 Workshop*.

[7] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1657–1664.

[8] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, Fabian Nater, and Luc Van Gool. 2013. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1633–1640.

[9] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, and others. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), e1390.

[10] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 1. IEEE, 419–426.

[11] Congcong Li and Tsuhan Chen. 2009. Aesthetic visual quality assessment of paintings. *IEEE Journal of selected topics in Signal Processing* 3, 2 (2009), 236–252.