LAPI at MediaEval 2016 Predicting Media Interestingness Task

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

ABSTRACT

This paper will present our results for the MediaEval 2016 Predicting Media Interestingness task. We proposed an approach based on video descriptors and studied several machine learning models, in order to detect the optimal configuration and combination for the descriptors and algorithms that compose our system.

1. INTRODUCTION

Interestingness is the ability to attract and hold human attention, this concept is gaining importance in the field of computer vision, especially since the growing importance and market value of social media and advertising. Even though the concept of interest might seem the result of a subjective viewer judgment, important progress has been made towards both an objective and context-based model for interest. Generally, in the field of computer vision two directions arose regarding this topic: pure visual interestingness (based on multimedia features and ideas [5, 6, 7]) and social interestingness (based on the degree of social media interest shown for certain visual data [5, 8]). Some researchers [8] focused on the similarities and differences between these two directions. Studies have been made regarding the psychological and physiological connections with novelty, enjoyment, challenge [1, 3], appraisal structures [10, 11] and computer vision concepts [5, 7, 6].

In this context, the MediaEval 2016 Predicting Media Interestingness Task [4] challenges the participants to automatically select images and/or video segments which are considered to be the most interesting for a common viewer. The concept of interestingness is defined in a particular use case scenario, i.e., helping professionals to illustrate a Video on Demand (VOD) web site by selecting some interesting frames and/or video excerpts for the movies. In this working note paper, we present our machine learning based approach to the task.

2. PROPOSED APPROACH

As previously stated, to determine the interestingness of images and video, we have experimented with a classic machine learning approach. First, the raw data is converted to content descriptors which should capture as best as possible the visual interestingness features of the data. Then, a supervised classifier is learned on these features using the labeled examples. Finally, the actual evaluation is carried out by feeding the classifier the unlabeled data. Regarding the content descriptors, we used the ones provided by the task organizers [4] with some additions. They were used as descriptors for a learning system based on SVM, where we tested different combinations of SVM kernel types and coefficients by using the LibSVM library [2].

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

2.1 Used features

Several visual features were used as descriptors, many of them being used in the literature for some computer vision tasks. The provided computed features were: color histogram of the Hue-Saturation-Value (denoted histo), Histogram of Oriented Gradients (HoG) descriptors computed over densely sampled patches, dense Scale Invariant Feature Transform (SIFT) with a codebook of 300 codewords and a three layered spatial pyramid (denoted dsift), Local Binary Patterns (LBP), GIST computed with the output of Gabor-like features (denoted gist) and the fc7 and prob layers of AlexNet (denoted cnnfc7 and cnnprob). All these features are presented and detailed in [4] and [9]. We also extracted and used the color naming histogram (denoted colornames) feature based on the work [12], as we wanted to obtain a color descriptor with fewer dimensions for our learning algorithms, that could better represent a human-centered understanding of the colors in each image or video.

For the *image subtask*, each image is represented with a content descriptor. For the *video subtask*, each video contains a certain number of images. To determine the final descriptor we use the simple averaging of the frames descriptors, leading in the end to a global descriptor per video.

2.2 Learning system

The learning is achieved using a Support Vector Machine (SVM) binary classifier. For all trained SVM models we used polynomial, RBF and linear kernels. For the polynomial kernels we used all the combinations of the following degrees : 1, 2, 3*k where $k \in [1, ..., 10]$ and the gamma coefficients were set as 2^k where $k \in [0, ..., 6]$. For the RBF kernel combinations we had values for the cost parameter of 2^k where $k \in [-4, ..., 8]$ and gamma coefficients with values in 2^k where $k \in [-4, ..., 8]$. We also tried different weights, considering the fact that the *devset* data, both for images and for videos, was unbalanced, the ratio of uninteresting to interesting samples being almost 10 to 1.

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

Subtask	Feature	SVM type	Degree	Gamma	TP	FP	Precision	Recall	MAP
image	histo+gist	poly	18	2	22	76	**0.224**	0.05	**0.214**
image	dsift+gist	poly	3	32	63	330	0.16	0.144	0.211
image	histo+dsift+gist	poly	9	2	15	35	0.3	0.034	0.197
image	colornames+any	poly	3	2	56	334	0.143	0.128	0.195
image	colornames	poly	2	8	226	1892	0.107	**0.517**	0.195
video	gist+cnnprob	poly	9	4	35	305	0.103	0.083	**0.179**
video	cnnfc7+any	poly	3	4	40	364	0.099	0.095	0.172
video	dsift+cnnprob	poly	24	64	81	846	0.087	**0.192**	0.159
video	gist	poly	6	8	49	359	**0.121**	0.116	0.148
video	dsift	poly	3	64	25	204	0.109	0.059	0.147

Table 2: Final results on *testset* (best results are marked in bold)

Run	Subtask	Feature	SVM Type	Degree	Gamma	MAP	P@5	P@10	P@20	P@100
run1	image	histo+gist	poly	18	2	**0.1714**	0.1077	0.1346	0.1423	0.0869
run2	image	dsift+gist	poly	3	32	0.1398	0.0462	0.0808	0.1000	0.0862
run3	video	gist+cnnprob	poly	9	4	0.1574	0.0923	0.1269	0.1212	0.0812
run4	video	cnnfc7+histo	poly	3	4	0.1572	0.1231	0.1000	0.1077	0.0815
run5	video	dsift+cnnprob	poly	24	64	**0.1629**	0.1154	0.1500	0.1192	0.0819

3. EXPERIMENTAL RESULTS

The task data consists of a development data intended to train the approaches and a test data for the actual benchmarking. The *devset* was extracted from 52 trailers, manually segmented, thus obtaining 5054 segments. For the *image subtask* one key-frame was used from each segment, while for the *video subtask* the whole segment was used. By annotating all the data a total of 473 interesting images and 420 interesting videos were obtained, with a provided interestingness score for calculating the mean average precision. The *testset* consisted of 26 trailers divided into 2342 segments. We performed a number of experiments on *devset* and selected the best combinations to be run on testset.

3.1 Experiments on devset

Using a 10-fold cross-validation, we chose the best results for the descriptor-classifier combinations based on precision, with a recall better than 0.03. For those best combinations we calculated the mean average precision. We have experimented with many different combinations of descriptors and SVM kernels. The best performing combination was generally the polynomial SVM. A high number of training runs, especially with the RBF or linear kernels, tended to classify all or almost all (low recall) the samples as non-interesting. In the case of weight-based training for the RBF kernel the recall tended to grow, but the precision was below that of the polynomial SVMs.

Table 1 lists the best five results for each of the two subtasks, giving details regarding the best coefficient combination used. As shown, the estimated MAP on the *devset* was better for the *image subtask* than for the *video subtask*. The MAP scores were calculated by using LibSVM's decision_values/prob_estimates output result for indicating the interestingness score of each sample [2]. The values for true positives, false positives, precision and recall are also listed. The best results were achieved with a descriptor composed of HSV Histogram and GIST, with a polynomial SVM with 18 degree and 2 gamma for the *image subtask*, and a descriptor composed of GIST and CNNProb layer, with a polynomial SVM with 9 degree and 4 gamma for the *video subtask*.

3.2 Official results on testset

The teams were allowed to submit 5 runs, so we chose the best 2 descriptor-classifier combinations for the *image subtask* and the best 3 combinations for the *video subtask*. This time the training of the SVM learning systems was done on the entire *devset*, using the optimal degree and gamma parameters obtained in our previous experiments. The submitted runs were the following : run1 - image subtask with HSV Histogram + GIST, SVM with degree = 18 and gamma = 2, run2 - image subtask with DSIFT + GIST, SVM with degree = 3 and gamma = 32, run3 - video subtask with GIST + CNNProb, SVM with degree = 9 and gamma = 4, run4 - video with CNNFc7 + HSV Histogram, SVM with degree = 3 and gamma = 4 and run5 - video with DSIFT + CNNProb, SVM with degree = 24 and gamma = 64.

The final results, as returned by the task organizers are presented in Table 2. The best results were a 0.1714 MAP on run1 for the *image subtask* and a 0.1629 MAP on run5 for the *video subtask*. With the single exception being run5, the MAP results on *testset* were below the estimated MAP on *devset*.

4. CONCLUSIONS

In this paper we presented several models for predicting and scoring multimedia interestingness. Our best MAP results on the *testset* were 0.1714 for the *image subtask* and 0.1629 for the *video subtask*. These results seem to indicate that the task in very challenging, one possible reason for this being the subjective nature of this field of study.

5. REFERENCES

[1] D. E. Berlyne. Conflict, arousal, and curiosity. 1960.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on*

Intelligent Systems and Technology (TIST), 2(3):27, 2011.

[3] A. Chen, P. W. Darst, and R. P. Pangrazi. An examination of situational interest and its sources. *British Journal of Educational Psychology*, 71(3):383–400, 2001.

[4] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. K. Duong, and F. Lefèbvre. Mediaeval 2016 predicting media interestingness task. In *Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands, Oct. 20-21, 2016.*

[5] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011.

[6] H. Grabner, F. Nater, M. Druey, and L. V. Gool. Visual interestingness in image sequences. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 1017–1026. ACM, 2013.

[7] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Gool. The interestingness of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1633–1640, 2013.

[8] L.-C. Hsieh, W. H. Hsu, and H.-C. Wang. Investigating and predicting social and visual image interestingness on social media by crowdsourcing. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4309–4313. IEEE, 2014.

[9] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang. Super fast event recognition in internet videos. *IEEE Transactions on Multimedia*, 17(8):1174–1186, 2015.

[10] P. J. Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89, 2005.

[11] S. A. Turner and P. J. Silvia. Must interesting things be pleasant? a test of competing appraisal structures. *Emotion*, 6(4):670, 2006.

[12] J. V. D. Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, 2009.