TCNJ-CS @ MediaEval 2017 Predicting Media Interestingness Task

Sejong Yoon
The College of New Jersey, USA
yoons@tcnj.edu

ABSTRACT

In this paper, we present our approach and investigation on the MedialEval 2017 Predicting Media Interestingness Task. We used most of the visual and audiotory features provided. The standard kernel fusion technique was applied to combine features and we used the ranking support vector machine to learn the classification model. No extra data was introduced to train the model. Official results, as well as our investigation on the task data is provided at the end.

1 INTRODUCTION

MediaEval 2017 Predicting Media Interestingness [2] consists of two subtasks. In the first task, the system should predict whether the viewer will consider a given image to be interesting or not to the common viewers. In the second task, a similar task should be performed given a video segment. In both tasks, the system should predict both the binary decision whether the media is interesting or not, and the ranking of the image frame/video segment among all image frame/video segments within the same movie. The data consists of 108 video clips. In total 7,396 key-frames and the same number of video segments are provided in the development set, and 2,436 key-frames and the same number of video segments are reserved for the test set. In this work, we used most of the features provided by the task organizers and we did not introduce any external data, e.g., meta-data, rating, reviews of the movies.

2 APPROACH

In this section, we first describe the features we employed and then present our classification method.

2.1 Features

We used features from different modalities. All features were provided by the task organizers.

Visual Features We used nearly all features provided, including Color histogram in HSV space, GIST [9], Dense SIFT [7], HOG 2x2 [1], Linear Binary Pattern (LBP) [8], prob (fc8, probabilities of predicted labels of 1,000 objects) layer of AlexNet [5], and C3D [10].

Audio Features We used the provided Mel-frequency Cepstral Coefficients (MFCC) features. An MFCC descriptor (60 dimensions) is computed over every 32ms temporal window with 16ms shift. The first and second derivatives of the cepstral vectors are also included in the MFCC descriptors.

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

For the image prediction task, we vectorized each feature per frame. For the video prediction task, we took the mean of raw feature values of all frames in the segment. Given the original feature $\mathbf{f}_{t,n}$ for n-th frame in t-th segment, we compute the summarized feature for the segment t as

$$\mathbf{x}_t = \frac{1}{N}\sum_{n=1}^{N}\mathbf{f}_{t,n} \tag{1}$$

where N denotes the total number of frames in the segment.

We used prob (fc8) layer to incorporate semantic information of the training data that can be extracted from the deep neural network.

2.2 Classification

We applied the standard kernel fusion approach: we compute a kernel for each type of features, and combine the kernels either by additions or multiplications. We used the multiplication within the same modality and we used the addition across the different modalities. For the image prediction subtask, we used the following combination of kernels:

$$K_1 = K_{chist} \cdot K_{gist}, \tag{2}$$

$$K_2 = K_{dhist} \cdot K_{hog} \cdot K_{lbp}, \tag{3}$$

$$K_{all} = K_1 + K_2 + K_{prob}. \tag{4}$$

The rational behind this choice was to consider features with global histograms and features using the spatial pyramids [6] as different modalities. We present the results on different kernel combinations for development set in the following section. The CNN probability layer, K_{prob}, is also considered as another modality since it conveys semantic information (objects in the images). For the video prediction subtask, we used the following combination of kernels:

$$K_{all} = K_1 + K_2 + K_{prob} + K_{c3d} + K_{mfcc}. \tag{5}$$

Since C3D and MFCC features model temporal aspect of input, we consider them as different modalities from the visual features. For the kernel choice, we used RBF kernel with the median of training data for the hyper-parameter choice.

For the classification model, we used the ranking support vector machine. We used SVM^{rank} [4] to learn pair-wise ranking patterns from the development set data, following a prior work [3].

3 RESULTS AND ANALYSIS

The official metric for evaluation is the mean average precision at 10 (MAP@10) computed over all videos, and over the top 10 best ranked images/video segments. First, we present different kernel combinations we tested on the development set. Table 1 describes the different kernel fusion formula we used in the experiments. We report both MAP and MAP@10 results in Table 2. As one can see,

Table 1: Different visual feature combinations

Combined kernel	Fusion formula
V_1	$K_1 \cdot K_2 \cdot K_{prob}$
V_2	$K_1 \cdot K_2 + K_{prob}$
V_3	$K_1 + K_2 + K_{prob}$

Table 2: Result of all subtasks in development set

Subtask	Measure	Result	Kernel
Image	MAP	0.3065	V_1
	MAP@10	0.0123	V_1
Image	MAP	0.3013	V_2
	MAP@10	0.0094	V_2
Image	MAP	0.3003	V_3
	MAP@10	0.0074	V_3
Video	MAP	0.3052	$V_1 + K_{c3d} + K_{mfcc}$
	MAP@10	0.0084	$V_1 + K_{c3d} + K_{mfcc}$
Video	MAP	0.3055	$V_2 + K_{c3d} + K_{mfcc}$
	MAP@10	0.0082	$V_2 + K_{c3d} + K_{mfcc}$
Video	MAP	0.3038	$V_3 + K_{c3d} + K_{mfcc}$
	MAP@10	0.0082	$V_3 + K_{c3d} + K_{mfcc}$

there is no significant differences among kernel fusion choices. We used 50-50 split, i.e. 39 movies each for train and test splits of the development set.

We also report both MAP and MAP@10 results on the testset in Table 3 provided by the task organizers. As described in the previous section, we used the visual feature combination, Eq. 4 for the image prediction task, and we used the multi-modal combination, Eq. 5 for the video prediction task. SVM^{rank} takes the ranking information as the label of input data and generates pairwise constraints. All provided ranking information in the development set was used for training the SVM^{rank} model, with image snapshots and video segments in each movie grouped together.

As it can be seen, in both image and video subtasks, the system shows low performance. This is not surprising given the very simple nature of the approach we applied to the task. What was not expected is that the video prediction result is much better (although still not reaching the level of good performance) than the image prediction result, which was not observable in the development set. This is interesting because we used the same set of features for image and video prediction subtasks, and the only differences are the two additional features modeling the temporal aspect of data (C3D, MFCC). We believe this reiterates a known understanding on the task: we must somehow incorporate temporal information to improve video interestingness prediction.

4 DISCUSSION AND OUTLOOK

One of the major challenges in video interestingness prediction is to fill the semantic gap. Initially, we intended to fill this gap by capturing expected emotional status of viewers and connect it to the notion of interestingness. Table 4 shows our categorization

Table 3: Result of all subtasks in testset

Subtask	Measure	Result	Kernel
Image	MAP	0.1331	V_3
	MAP@10	0.0126	V_3
Video	MAP	0.1774	$V_3 + K_{c3d} + K_{mfcc}$
	MAP@10	0.0524	$V_3 + K_{c3d} + K_{mfcc}$

Table 4: Key-frames of most interesting segments in some development set movies categorized into types of interest stimuli

Subtask	Key-frames
Violence	
Nudity	
Horror / Surprise	
Romantic mood	
Facial expression	
Joyful, Fun, Humor	
Open view / scenery	
Others (context)	

of the most interesting segments in each movie clip we gathered during the progress. As it can be seen, many of the categories are closely related to key emotional states that modern and existing affect prediction methods can predict. This is particularly true for violence, horror, and joy which consist in large proportion of the most interesting video segments. On the other hand, there are many other video segments for which one cannot readily identify the root of interest stimuli. These typically require a higher level of understanding of the context. The best example is the third movie in the Others category which requires fusion of all modalities plus reading of a sentence shown on the image frame.

In the future, we hope to challenge the media interestingness prediction problem in this direction. Maybe the most promising approach at this point is to understand human activities and link them to emotions and the interestingness.

ACKNOWLEDGMENTS

This work was supported in part by The College of New Jersey under Support Of Scholarly Activity (SOSA) 2017-2019 grant.

REFERENCES

[1] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 886–893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[2] Claire-Hélène Demarty, Mats Sjöberg, Bogdan Ionescu, Thanh-Toan Do, Michael Gygli, and Ngoc Q. K. Duong. Predicting Media Interestingness Task at MediaEval 2017. In *Proc. of MediaEval 2017 Workshop, Dublin, Ireland, Sept. 13-15, 2017.*

[3] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. 2013. Understanding and Predicting Interestingness of Videos. In *AAAI.*

[4] Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).*

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25 (NIPS).*

[6] S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 2. 2169–2178. https://doi.org/10.1109/CVPR.2006.68

[7] David Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision* (2004).

[8] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. 2002. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 7 (July 2002), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623

[9] Aude Oliva and Antonio Torralba. 2001. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int. J. Comput. Vision* 42, 3 (May 2001), 145–175. https://doi.org/10.1023/A:1011139631724

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV.*