NII-UIT at MediaEval 2016 Predicting Media Interestingness Task

Vu Lam
University of Science,
VNU-HCM
lqvu@fit.hcmus.edu.vn

Tien Do
University of Information
Technology, VNU-HCM
tiendv@uit.edu.vn

Sang Phan
National Institute of
Informatics, Japan
plsang@nii.ac.jp

Duy-Dinh Le
National Institute of
Informatics, Japan
ledduy@nii.ac.jp

Shin'ichi Satoh
National Institute of
Informatics, Japan
satoh@nii.ac.jp

Duc Anh Duong
University of Information
Technology, VNU-HCM
ducda@uit.edu.vn

ABSTRACT

The MediaEval 2016 Predicting Media Interestingness (PMI) Task requires participants to retrieve images and video segments that are considered to be the most interesting for a common viewer. This is a challenging problem not only because the large complexity of the data but also due to the semantic meaning of interestingness. This paper provides an overview of our framework used in MediaEval 2016 for the PMI task and discusses the performance results for both subtasks of predicting image and video interestingness. Experimental results show that, our framework give a reasonable accuracy just by simply using low-level features: GIST, HoG, Dense SIFT, and incorporating deep features from pre-trained deep learning models.

1. INTRODUCTION

Following the setting of this task [3], we design a framework that consists of three main components: feature extraction and encoding, feature classification, and feature fusion. An overview of our framework is shown in Fig 1. For the features extracted from video frames, we use the max pooling strategy to aggregate all frame features of a same shot to form the shot representation. In the training step, we train a classifier for each type of features using the Support Vector Machine [1]. Then we use these classifiers to predict the scores for each shot. Finally, we adopt the late fusion with average weighting scheme to combine the prediction scores of various features.

2. FEATURE EXTRACTION

2.1 Low-level Features

We use features that are provided by the organizers [6]. More specifically, following features are exploited for the task.

- **Dense SIFT** are computed following the original work in [9], except that the local frame patches are densely sampled instead of using interest point detectors. A codebook of 300 code words is used in the quantization process with a spatial pyramid of three layers [8];
- **HOG** descriptors [2] are computed over densely sampled patches. Following [12], HOG descriptors in a 2x2 neighborhood are concatenated to form a descriptor of higher dimension;
- **GIST** is computed based on the output energy of several Gabor-like filters (8 orientations and 4 scales) over a dense frame grid like in [10].

2.2 Audio Features

In predicting video interestingness task, we use the popular Mel-frequency Cepstral Coefficients (MFCC) for extracting audio features. We choose a length of 25ms for audio segments and a step size of 10ms. The 13-dimensional MFCC vectors along with each first and second derivatives are used for representing each audio segment. Raw MFCC features are also encoded using Fisher vector encoding. We use a GMM to train the codebook with 256 clusters. For audio features, we do not use PCA. The final feature descriptor has 19,968 dimensions.

2.3 Deep Features

We used the popular Caffe framework [5] to extract deep features from two pre-trained model Alexnet [7] and VGG [11]. These models were trained on ImageNet 1,000 concepts [4].

AlexNet is the first work that popularized Convolutional Networks in Computer Vision, developed by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton. It is the winning system of ILSVRC2012 classification task [4] and it outperformed other methods by a large margin in terms of accuracy. This very first visual deep learning network only contains 5 convolutional layers and 3 fully-connected layers.

VGGNet refers to a deep convolutional network for object recognition developed and trained by Oxford's renowned Visual Geometry Group [11]. They provided two deep networks that consist of 16 and 19 layers respectively. In our experiments, we use the VGGNet with 16 layers for feature extraction.

We selected the neuron activations from the last three layers for the feature representation. The third and second-to-last layer has 4,096 dimensions, while the last layer has 1,000 dimensions corresponding to the 1,000 concept categories in the ImageNet dataset. We denote these features

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

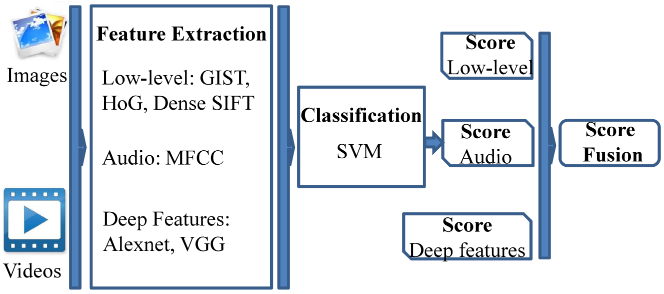

Figure 1: Our framework for extracting and encoding local features.

Table 1: Results of predicting interestingness from image

Run	Features	Results (MAP)
FA	VGGFC8+AlexNetFC8	21.15
V1	VGGFC7+GIST+HOG+ DenseSIFT	17.73

as AlexNetFC6, AlexNetFC7, AlexNetF8, VGGFC6, VGGFC7, and VGGFC8 in our experiments.

3. CLASSIFICATION

LibSVM [1] is used for training and testing our interestingness classifiers. For features that are encoded using the Fisher vector, we use linear kernel for training and testing. For deep learning feature, χ_2 kernel is used. The optimal gamma and cost parameters for learning SVM classifiers are found by conducting a grid search with 5-fold cross validation on the training dataset.

4. SUBMITTED RUNS

At first, we use the late fusion with average weighting scheme to combine features from different modalities. After that we select the runs that have the top performance on the validation set to submit. The list of submitted runs for each subtask and its results can be seen on Table 1 and Table 2.

5. RESULTS AND DISCUSSIONS

The official results for each subtask are shown on the last column of Table 1 and Table 2, which are corresponding to the results of predicting interestingness from image and video respectively. These results show that predicting interestingness from image is more accurate than from video. This can be due to the highly dynamic of video content. Moreover, the performance of predicting interestingness from video can be improved if motion features are exploited, which have not been incorporated to our system for the time being.

Table 2: Results of predicting interestingness from video

Run	Features	Results (MAP)
FA	AlexNetFC8+MFCC	16.9
F1	VGGFC7 + GIST	16.41

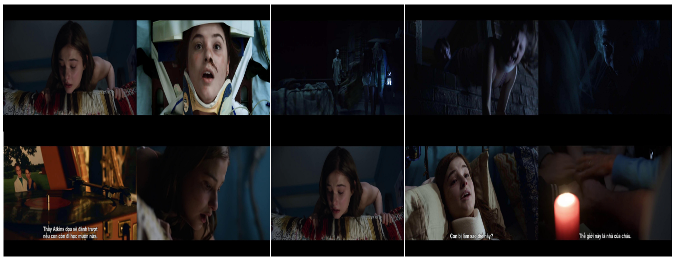

Figure 2: Top interesting images of detected by our system.

Examples of top interesting images that are detected by our system are illustrated on Fig. 2. Interestingly, our system tends to output a higher rank on images of beautiful women. Furthermore, we found that images from dark scenes are often considered more interesting, probably because these scenes often draw more attention from the audiences.

6. ACKNOWLEDGEMENTS

This research is partially funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2013-26-01.

7. REFERENCES

[1] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[3] C.-H. Demarty, M. Sjöberg, B. Ionescu, T.-T. Do, H. Wang, N. Q. Duong, and F. Lefebvre. Mediaeval 2016 predicting media interestingnesstask. *Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands, Oct. 20-21, 2016.*

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[6] 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.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] S. Lazebnik, C. Schmid, and J. Ponce. 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)*, volume 2, pages 2169–2178. IEEE, 2006.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[10] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.