Ranking Images and Videos on Visual Interestingness by Visual Sentiment Features

Soheil Rayatdoost
Swiss Center for Affective Sciences
University of Geneva
Switzerland
soheil.rayatdoost@unige.ch

Mohammad Soleymani
Swiss Center for Affective Sciences
University of Geneva
Switzerland
mohammad.soleymani@unige.ch

ABSTRACT

Today, users generate and consume millions of videos online. Automatic identification of the most interesting moments of these videos have many applications such as video retrieval. Although most interesting excerpts are person-dependent, existing work demonstrate that there are some common features among these segments. The media interestingness task at MediaEval 2016 focuses on ranking the shots and key-frames in a movie trailer based on their interestingness. The dataset consists of a set of commercial movie trailers from which the participants are required to automatically identify the most interesting shots and frames. We approach the problem as a regression task and test several algorithms. We particularly use mid-level semantic visual sentiment features. These features are related to the emotional content of images and are shown to be effective in recognizing interestingness in GIFs. We found that our suggested features outperform the baseline for the task at hand.

1. INTRODUCTION

Interestingness is the capability of catching and holding human attention [1]. Research in psychology suggests that interest is related to novelty, uncertainty, conflict and complexity [2, 14]. These attributes determine whether a person finds an item interesting. The attributes contribute to interestingness differently for different people, for example, one might find more complex stimulus more interesting than the other. Developing a computational model which automatically perform such a task is useful for different applications such as video retrieval, recommendation and summarization [1, 15].

There are a number of work that address the problem of visual interestingness prediction from the content. Gygli et al. and Grabner et al. [7, 6] used visual content features related to unusualness, aesthetics and general preference for predicting visual interestingness. Soleymani [15] built a model for personalized interest prediction for images. He found that affective content, quality, coping potential and complexity have a significant effect on visual interest in images. In a more recent work, Gygli and Soleymani [8] attempted predicting GIF interestingness from the content. They found that visual sentiment descriptors [11] to be more effective for predicting GIF interestingness compared to the features that capture temporal information and motion.

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

Figure 1: Examples of hit (top row) and miss (bottom row) top-ranking key-frames.

The "media interestingness task" is organized at MediaEval 2016. In this task, a development and evaluation-set consisting of Creative Commons licensed trailers of commercial movies with their interestingness labels are provided. For the details of the task description, dataset development and evaluation, we refer the reader to the task overview paper [3]. There are two subtasks for this challenge, the first one involves automatic prediction of interestingness ranking for different shots in a trailer. The second task involves predicting the ranking for the most interesting key frames. Visual and audio (only for shots) modalities are available for the interestingness prediction methods [3]. The designed algorithms are evaluated over evaluation data which include 2342 shots from 26 trailers. Examples of top-ranking key-frames are shown in Figure 1.

The organizers provided a set of baseline visual and audio features. For the visual modality, we additionally extracted mid-level semantic visual descriptors [11] and deep learning features. Sentiment related features are effective in capturing emotional content of images and are shown to be useful in recognizing interestingness in GIFs [8]. For the audio modality, we extracted the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [4]. We tested multiple regression models for interestingness ranking. We compare our results with the ones from the baseline features based on mean average precision (MAP) over top N best ranked images or shots. According to our results on the evaluation-set, our feature-set outperform the baseline features for predicting interestingness. In the next section, we present our features and describe our methodology in detail.

2. METHOD

2.1 Features

We opt for using a set of hand-crafted features and transfer learning in addition to regression models with the goal of interestingness ranking. The task organizers provided a set of baseline low-level features. These features include a number of low-level audiovisual features that are typically used for computer vision and speech analysis, including dense SIFT, Histogram of Gradients (HoG), Local Binary Patterns (LBP), GIST, Color Histogram, deep learning features for the visual modality [10], and Mel-Frequency Cepstral Coefficients (MFCC) and the cepstral vectors for audio.

Interestingness is highly correlated with image emotional content [15]. Therefore, we opted for extracting the eGeMAPS from audio [4]. eGeMAPS features are acoustic features hand-picked by experts for the goal of speech and music emotion recognition. 88 eGeMAPS features were extracted by openSMILE [5]. For video sub-challenge, we extracted all the key-frames from each shot. We then applied the visual sentiment adjective-noun-pair (ANP) detectors [11] on each key-frame. The weights from the fully connected layer 7 (fc7) and the output from the final layer was extracted on each frame. We then pooled the resulting values by mean and variance to form one feature vector for each shot.

2.2 Regression models

We used three different regression models to predict the interestingness level (linear regression (LR), support vector regression (SVR) with linear kernel and sparse approximation weighted regression (SPARROW) [13].

We used LIBLINEAR Library [9, 12] implementation of SVR with L2-regularized logistic regression option to predict the interesingness score. We also used a regression with sparse approximation. Regression with sparse approximation is a regression model for approximation of the prediction based on local information. It is similar to a k-nearest neighbors regression (k-NNR) whose weights are calculated based on sparse approximation [13]. Linear regression with minimum least-squares optimization is utilized as a baseline method.

In all cases, except eGeMAPS audio features, we used principal component analysis (PCA) to reduce the dimensionality of features. For SVR and SPARROW, we kept the principal components containing 99% of variance. In case of linear regression, we only kept the principal components that added up to 50% of the total variance.

3. EXPERIMENTS

After extracting all the feature-sets, we evaluated the performance of different combinations of the feature-sets and regression models. We evaluated different approaches using a five-folding cross-validation on the development-set. In each iteration, one-fifth of the development-set was held out and the rest was used to train the regression model. When training the SVR, we optimized the hyper-parameter C using a grid-search on the training-set.

The best performing approaches based on their performance measured by MAP on the ranked results were selected for submitted runs (See Table 1).

4. RESULTS AND DISCUSSION

Following the task evaluation procedure, we report MAP on N best ranked images or shots. We report the results on the cross-validation on the development-set and on our four submitted runs on the evaluation-set. For our submitted runs, we trained selected features and regression methods on all the available data in the development-set. The results for interestingness prediction with the best pair of regression methods and feature-sets are summarized in Table 1. The best MAP on the development-set which is achieved by combining multilingual visual sentiment ontology (MVSO) descriptors and deep learning features in combination with SPARROW regression is 0.262. We used Baseline video features and SPARROW regression as our baseline. To check the performance of audio features we ranked the video with respect to SVR output which was trained on audio features only. The best results for image sub-task is achieved by sentiment descriptors and deep learning features in combination with linear regression.

Table 1: Evaluation results on interestingness ranking.

	Task	Method	Features	MAP ↑
Dev. Set	Image	LR	MVSO+fc7	0.1710
	Video	SPARROW	MVSO+fc7	0.2617
	Video	SPARROW	Baseline	0.2414
	Video	SVR	eGeMAPS	0.1987
Eval. Set	Image	LR	MVSO+fc7	0.1704
	Video	SPARROW	MVSO+fc7	0.1710
	Video	SPARROW	Baseline	0.1497
	Video	SVR	eGeMAPS	0.1367

Overall, the evaluation-set results demonstrate that mid-level semantic visual descriptors are more effective in predicting interestingness compared to the baselines low-level features. The results from a set of relatively simple audio features show the significance of audio modality for such a task. In Image sub-task, the evaluation-set results are very similar to video sub-task, since sentiment features lack temporal information. The drop in the performance on the evaluation-set demonstrates that our models were overfitting to the development-set and it is likely that an ensemble learning regression would have performed better.

5. CONCLUSION

In this work, we explored different strategies for predicting visual interestingness in videos. We found the mid-level visual descriptors which are related to sentiment to be more effective for such a task compared to the low-level visual features. This is due to the affective nature of interestingness, i.e., interest is an emotion by some account. Our features are all static and frame-based; we did not try extracting features related to movement that can capture temporal information due to the small size of the dataset. Hence, the frame-based results are not any different to the shot-based ones. Essentially they do very similar tasks. The observed performance of the proposed method is rather low. However given the sample size and the dimensionality of the descriptors, they still show promising potential. In the future, ideally larger scale datasets shall be developed and annotated to enable using more sophisticated methods such as transfer learning using deep neural networks. Even though the audio features are not as effective, they showed significant performance deserving more in-depth analysis in the future.

6. REFERENCES

[1] X. Amengual, A. Bosch, and J. L. de la Rosa. Review of methods to predict social image interestingness and memorability. In G. Azzopardi and N. Petkov, editors, *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I*, pages 64–76. Springer International Publishing, Cham, 2015.

[2] D. Berlyne. *Conflict, arousal, and curiosity.* McGraw-Hill, 1960.

[3] C. Demarty, M. Sjöberg, B. Ionescu, T. Do, H. Wang, N. Duong, and F. Lefebvre. Mediaeval 2016 predicting media interestingness task. *In MediaEval 2016 workshop, Amsterdam, Netherland*, 2016.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, April 2016.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM.

[6] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *Proceedings of the 21st Annual ACM Conference on Multimedia*, 2013.

[7] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The Interestingness of Images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2013.

[8] M. Gygli and M. Soleymani. Analyzing and predicting GIF interestingness. In *ACM Multimedia*, 2016.

[9] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the Twenty Fifth International Conference on Machine Learning (ICML)*, 2008.

[10] 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, Aug 2015.

[11] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 159–168, New York, NY, USA, 2015. ACM.

[12] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region Newton method for large-scale logistic regression. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007.

[13] P. Noorzad and B. L. Sturm. Regression with sparse approximations of data. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 674–678, Aug 2012.

[14] P. J. Silvia, R. A. Henson, and J. L. Templin. Are the sources of interest the same for everyone? using multilevel mixture models to explore individual differences in appraisal structures. *Cognition and Emotion*, 23(7):1389–1406, 2009.

[15] M. Soleymani. The quest for visual interest. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 919–922, New York, NY, USA, 2015. ACM.