The ICL-TUM-PASSAU Approach for the MediaEval 2015 “Affective Impact of Movies” Task

George Trigeorgis[1], Eduardo Coutinho[1],
Fabien Ringeval[2,3], Erik Marchi[2],
Stefanos Zafeiriou[1], Björn Schuller[1,3]
[1]Department of Computing, Imperial College London, UK
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Munich, Germany
[3]Chair of Complex & Intelligent Systems, University of Passau, Germany
g.trigeorgis@imperial.ac.uk

ABSTRACT

In this paper we describe the Imperial College London, Technische Universität München and University of Passau (ICL+TUM+PASSAU) team approach to the MediaEval's “Affective Impact of Movies” challenge, which consists in the automatic detection of affective (arousal and valence) and violent content in movie excerpts. In addition to the baseline features, we computed spectral and energy related acoustic features, and the probability of various objects being present in the video. Random Forests, AdaBoost and Support Vector Machines were used as classification methods. Best results show that the dataset is highly challenging for both affect and violence detection tasks, mainly because of issues in inter-rater agreement and data scarcity.

1. INTRODUCTION

The MediaEval 2015 Challenge “Affective Impact of Movies” comprises two subtasks using the LIRIS-ACCEDE database [2]. *Subtask 1* targets the automatic categorisation of videos in terms of their affective impact. The goal is to identify the arousal (calm-neutral-excited) and valence (negative-neutral-positive) levels of each video. The goal of *Subtask 2* is to identify those videos that contain violent scenes. The full description of the tasks can be found in [22].

2. METHODOLOGY

2.1 Subtask 1: affect classification

Feature sets.

In our work we have used both the baseline features provided by the organisers [2], as well as our own sets of audiovisual features as described below.

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was used to extract acoustic features with the openSMILE toolkit [6]; this feature set was designed as a standard acoustic parameter set for automatic speech emotion recognition [5, 18, 16] and has also been successfully used for other paralinguistic tasks [17]. The eGeMAPS comprises a total of 18 Low-Level Descriptors (LLDs), including frequency, energy/amplitude, and spectral related features. Various functionals were then applied to the LLDs over the whole instance, giving raise to a total of 88 features.

The emotional impact of videos can be heavily influenced by the kind of objects present in a given scene [11, 12, 15]. We thus computed a probability of 1000 different objects to be present in a frame using a pretrained 16-layer convolutional neural network (CNN) on the ILSVRC2013 dataset [21, 4]. Let $\mathbf{x} \in \mathbb{R}^{N\times p}$ represents a video of the database with N frames and p pixels per frame, and $f(\cdot)$ the trained convolutional neural net with softmax activation functions in the output layer. The probability $Pr(y = c|x_i;\theta)$ for each of the 1000 classes being present inside the i-th frame of a video $\mathbf{x}_i$ is obtained by forwarding the p pixels value through the network. By averaging the activations over all the N frames of a video sequence we obtained the probability distribution of the 1000 ILSVRC2013 classes that might be present in the video.

Classifiers.

For modelling the data we concentrated on two out-of-the-box ensemble techniques: Random Forests and AdaBoost. We used these two techniques as they are less susceptible to the overfitting problem than other learning algorithms due to the combination of weak learners, they are trivial to optimise as they have only one hyper-parameter, and they usually provide close or on par results with the state-of-the-art for a multitude of tasks [9, 10, 23, 14]. The hyper-parameters for each classifier were determined using a 5-fold cross-validation scheme on the development set. During development the best performance was achieved with 10 trees with Random Forests and 20 trees with AdaBoost.

Runs.

We submitted a total of five runs. Run 1 consisted of predictions using the baseline features and the AdaBoost model. The predictions in runs 2 and 5 were obtained using the baseline plus our audio-visual feature sets and the Random Forest and AdaBoost classifiers, respectively. By looking at the distribution of labels in the development set, we observed that the most common combinations of labels are: 1) neutral valence (V^n) and negative arousal (A^-) (24%), and 2) positive valence (V^+) and negative arousal (A^-) (20%). Runs 3 and 4 are thus based on the hypothesis that the label distribution of the test set will be similarly unbalanced. In run 3 every clip was predicted to be V^n, A^+ and in Run 4 every one was V^+, A^-. These submissions act as a sanity check of our own models, but also other competitors' submissions for this competition.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2.2 Subtask 2: violence detection

Feature sets.

According to previous work [7, 13], we only considered spectral and energy based features as acoustic descriptors. Indeed, violent segments do not necessarily contain speech; voice specific features, such as voice quality and pitch related descriptors, might thus not be a reliable source of information for violence. We extracted 22 acoustic low-level descriptors (LLDs): loudness, alpha ratio, Hammarberg's index, energy slope and proportion in the bands $[0-500]$ Hz and $[500-1500]$ Hz, and 14 MFCCs, using the openSMILE toolkit [6]. All LLDs, with the exception of loudness and the measures of energy proportion, were computed separately for voiced and unvoiced segments. As the frames of the movie that contain violent scenes are unknown, we computed 5 functionals (max, min, range, arithmetic mean and standard-deviation) to summarise the LLDs over the movie excerpt, which provided a total of 300 features. For the video modality, we used the same additional features defined in *Subtask 1*. We also used the metadata information of the video genre as an additional feature, due to dependencies between movie genre and violent content.

Classifier.

Since the dataset is strongly imbalanced – only 272 excerpts out of 6,144 are labelled as violent – we up-sampled the violent instances to achieve a balanced distribution. All features were furthermore standardised with a z-score. As classifier, we used the `libsvm` implementation of Support Vector Machines (SVMs) [3] and optimised the complexity parameter, and the γ coefficient of the radial basis kernel in a 5-folds cross-validation framework on the development set. Because the official scoring script requires the computation of *a posteriori* probabilities, which is more time consuming than the straightforward classification task, we optimised the Unweighted Average Recall (UAR) to find the best hyper-parameters [19, 20], and then re-trained the SVMs with the probability estimates.

Runs.

We first performed experiments with the full baseline feature set and found that the addition of the movie genre as feature improved the Mean Average Precision (MAP) from 19.5 to 20.3, despite degrading the UAR from 72.3 to 72.0. Adding our own audio-visual features provided a jump in the performance with the MAP reaching 33.6 and UAR 77.6. Because some movie excerpts contain partly relevant acoustic information, we empirically defined a threshold on loudness based on the histogram, to exclude frames before computing the functionals. This procedure has improved the MAP to 35.9 but downgraded the UAR to 76.9. A fine tuning of the complexity parameter and γ coefficient yielded the best performance in terms of UAR with a value of 78.0, but slightly deteriorated the MAP to 35.7.

We submitted a total of five runs. Run 1 – baseline features; Run 2 – all features mentioned above (except movie genre) with loudness threshold (0.038); Run 3 – same as Run 2 plus the inclusion of movie genre; Run 4 – as Run 3 but with a fine tuning of the hyper-parameters; Run 5 – similar to Run 3 but with a higher threshold for loudness (0.078).

3. RESULTS

Our official results on the test set for both subtasks are shown in Table 1.

Subtask 1. Our results for the affective task indicate that we did not do much better than was expected by chance for arousal classification, and did slightly better than chance for valence in run 5; we thus refrain from further interpretation of results. This can be explained by the low quality of the provided annotations for the dataset. The initial annotations had a low inter-rater agreement [2], and there were multiple processing stages afterwards [1, 22] with high levels of uncertainty and unclear validity.

Subtask 2. Results show that there is an important overfitting in our models as the performance is divided by a factor of 2 between development and test partitions. This is, however, not really surprising since only 272 instances labelled as violent were available as training data. Moreover, the labelling task being performed not at the frame level but rather at the excerpt level does not allow to model precisely the information that is judged as violent, making the task highly challenging. We can nevertheless observe that the proposed audio-visual feature set brings a large improvement over the baseline feature set – the MAP is improved by a factor superior to 2, and that the inclusion of the movie genre as additional feature also allows a small improvement in the performance.

	Subtask 1		Subtask 2
Run	Arousal (AC)	Valence (AC)	Violence (MAP)
1	55.72	39.99	4.9
2	54.71	41.00	13.3
3	55.55	37.87	13.5
4	55.55	29.02	14.9
5	54.46	41.48	13.9

Table 1: The submission results for the arousal, valence, and violence classification tasks on the test partition. AC stands for accuracy and MAP for the mean average precision.

4. CONCLUSIONS

We have presented our approach to the MediaEval's "Affective Impact of Movies" challenge, which consists in the automatic detection of affective and violent content in movie excerpts. Our results for the affective task have shown that we did not do much better than a classifier that is based on chance, although we use features and classifiers that are known to work well in the literature for arousal and valence prediction [2, 8]. We consider that this might be owed to a potentially noisiness of the annotations provided. As for the violence prediction subtask, the results show that we overfit a lot on the development set, which is not very striking given the small amount of instances of the minority class. The analysis of violent content at the excerpt level is also highly challenging, because only few frames might contain violence, and such brief information is almost totally lost in the computation of functionals at the full excerpt level.

5. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the EC's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the EU's Horizon 2020 Programme through the Innovative Action No. 644632 (MixedEmotions), No. 645094 (SEWA) and the Research Innovative Action No. 645378 (ARIA-VALUSPA).

6. REFERENCES

[1] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. From crowdsourced rankings to affective ratings. In *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW 2014)*, pages 1–6, 2014.

[2] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. LIRIS-ACCEDE: A Video Database for Affective Content Analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, January-March 2015.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), April 2011.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255, 2009.

[5] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, In press, 2015.

[6] 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 2013)*, pages 835–838, Barcelona, Spain, October 2013.

[7] F. Eyben, F. Weninger, N. Lehment, and B. Schuller. Affective Video Retrieval: Violence Detection in Hollywood Movies by Large-Scale Segmental Feature Extraction. *PLOS one*, 8(12):1–12, December 2013.

[8] K. Forbes-Riley and D. J. Litman. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL)*, pages 201–208, 2004.

[9] F. G., J. Gall, and V. G. L. Real time head pose estimation with random regression forests. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 617–624, Providence (RI), USA, June 2011.

[10] L. Guo, N. Chehata, C. Mallet, and S. Boukir. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(1):56–66, January 2011.

[11] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[12] W. Hu, N. Xie, L. Li, X. Zeng, and M. S. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 41(6):797–819, October 2011.

[13] B. Ionescu, J. Schlüter, I. Mironica, and M. Schedl. A naive mid-level concept-based fusion approach to violence detection in hollywood movies. In *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval (ICMR 2013)*, pages 215–222, Dallas (TX), USA, 2013.

[14] J.-J. Lee, P.-H. Lee, S.-W. Lee, A. Yuille, and C. Koch. Adaboost for text detection in natural scene. In *Proceedings of the IEEE 12th International Conference on Document Analysis and Recognition (ICDAR 2013)*, pages 429–434, Beijing, China, 2013.

[15] I. Lopatovskaa and I. Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*, 47(4):575–592, July 2011.

[16] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller. Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion. In *Proceedings of the 2nd Emotion Recognition In The Wild Challenge and Workshop (EmotiW 2014)*, pages 473–480, Istanbul, Turkey, September 2014.

[17] F. Ringeval, E. Marchi, M. Mehu, K. Scherer, and B. Schuller. Face reading from speech – predicting facial action units from audio cues. In *Proceedings of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*, to appear, Dresden, Germany, September 2015.

[18] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), ACM MM*, Brisbane, Australia, October 2015.

[19] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *Proceedings of INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association (ISCA)*, pages 427–431, Singapore, Republic of Singapore, September 2014.

[20] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition. In *Proceedings of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*, Dresden, Germany, September 2015.

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

[22] M. Sjöberg, Y. Baveye, H. Wang, V. Quand, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The MediaEval 2015 Affective Impact of Movies Task. In *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 2015.

[23] A. Stumpf and N. Kerle. Object-oriented mapping of landslides using random forests. *Remote Sensing of Environment*, 115(10):2564–2577, October 2011.