TCS-ILAB - MediaEval 2015: Affective Impact of Movies and Violent Scene Detection

Rupayan Chakraborty, Avinash Kumar Maurya, Meghna Pandharipande,
Ehtesham Hassan, Hiranmay Ghosh and Sunil Kumar Kopparapu
TCS Innovation Labs-Mumbai and Delhi, India
{rupayan.chakraborty, avinashkumar.maurya, meghna.pandharipande,
ehtesham.hassan, hiranmay.ghosh, sunilkumar.kopparapu}@tcs.com

ABSTRACT

This paper describes the participation of TCS-ILAB in the MediaEval 2015 Affective Impact of Movies Task (includes Violent Scene Detection). We propose to detect the affective impacts and the violent content in the video clips using two different classifications methodologies, i.e. Bayesian Network approach and Artificial Neural Network approach. Experiments with different combinations of features make up for the five run submissions.

1. SYSTEM DESCRIPTION

1.1 Bayesian network based valence, arousal and violence detection

We describe the use of a Bayesian network for the detection task of violence/non-violence, and induced affect. Here, we learn the relationship between different attributes of different types of features using a Bayesian network (BN). Individual attributes such as Colorfulness, Shot length, or Zero Crossing etc. form the nodes of BN. This includes the valence, arousal and violence labels which are treated as categorical attributes. The primary objective of the BN based approach is to discover the cause-effect relationship between different attributes which otherwise is difficult to learn using other learning methods. This analysis helps in gaining the knowledge of internal processes of feature generation with respect to the labels in question, i.e. violence, valence and arousal.

In this work, we have used a publicly available Bayesian network learner [1] which gives us the network structure describing dependencies between different attributes. Using the discovered structure, we compute the conditional probabilities for the root and its cause attributes. Further, we perform inferencing for valence, arousal and violence values for new observations using the junction-tree algorithm supported in Dlib-ml library [2].

As will be shown later, conditional probability computation is a relatively simple task for a network having few nodes which is the case for image features. However, as the attribute set grows, the number parameters namely, conditional probability tables grow exponentially. Considering that our major focus is on determining the values of violence, valence and arousal values with respect to unknown values of different features, we apply the D-separation principle [3] for recursive pruning the network as it is not necessary to propagate information along every path in the network. This reduces the computational complexity by a significant level both for parameter computation and inference. Also, with pruned network, we observe a reduced set of features which effect the values of the queried nodes.

1.2 Artificial neural network based valence, arousal and violence detection

This section describes the system that uses Aritificial Neural Networks (ANN) for classification. Two different methodologies are employed for the two different subtasks. For both subtasks, the developed systems extract the features from the video shots (including the audio) prior to classification.

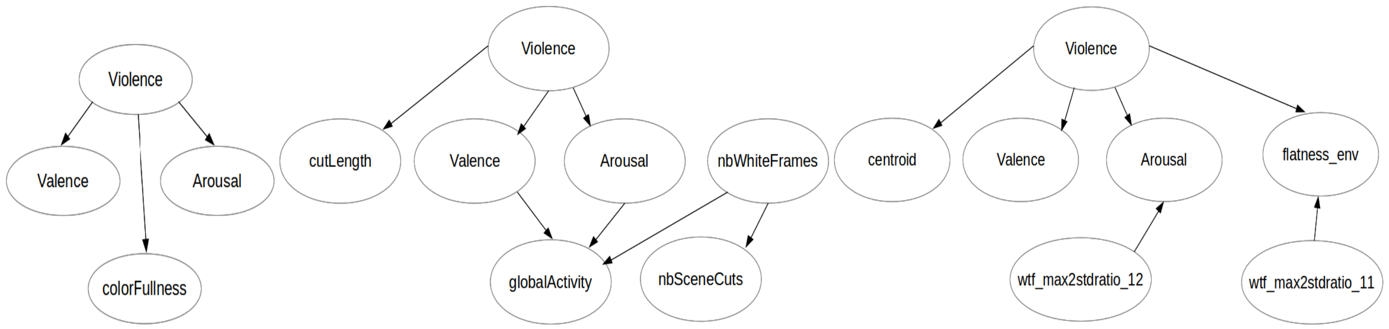
1.2.1 Feature extraction

The proposed system uses different set of features, either from the available feature set (audio, video, and image), which was provided with the MediaEval dataset, or from our own set of extracted audio features. The designed system either uses audio, image, video features separately, or a combination of them. The audio features are extracted using the openSMILE toolkit [4], from the audio extracted from the video shots. openSMILE uses low level descriptors (LLDs), followed by statistical functionals for extracting meaningful and informative set of audio features. The feature set contains the following LLDs: intensity, loudness, 12 MFCC, pitch (F0), voicing probability, F0 envelope, 8 LSF (Line Spectral Frequencies), zero-crossing rate. Delta regression coefficients are computed from these LLDs, and the following functionals are applied to the LLDs and the delta coefficients: maximum and minimum value and respective relative position within input, range, arithmetic mean, two linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile, and three inter-quartile ranges. openSMILE, in two different configurations, allows extractions of 988 and 384 (which was earlier used for Interspeech 2009 Emotion Challenge [5]) audio features. Both of these are reduced to a lower dimension after feature selection.

1.2.2 Classification

For classification, we have used an ANN that is trained with the development set samples available for each of those subtask. As data imbalance exists for the violence detection task (only 4.4% samples are violent), for training, we have taken equal number of samples from both the classes.

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



(a) Image features

(b) Video features

(c) Audio features

Figure 1: Pruned Bayesian Networks.

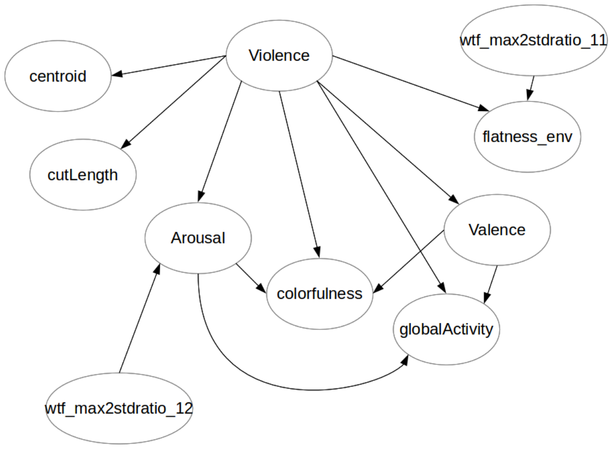


Figure 2: Pruned Bayesian Network with combined features.

Therefore, we have multiple ANNs, each of them is trained with different set of data. While testing, the test sample is fed to all ANNs, and then the scores from all ANNs outputs are summed using an add rule of combination and the class that has maximum score is declared winner alogwith a confidence score. Moreover, while working with the test dataset, the above mentioned framework is used with different feature sets. For combining the output of ANNs, two different methodologies are adopted. In the first, all the scores are added using an add rule before deciding on the detected class. In the second, the best neural network (selection based on development set) is used for each feature set. Finally, the scores from all the best networks are summed and the decision is made on the maximum score.

2. EXPERIMENTS

The BN network is learned using only the features provided with the MediaEval's 2015 development set [6] [7]. As given, the violence, valence, and arousal are categorical attributes where violence is a binary variable, and valence and arousal are distributed in three discrete states. For computing the prior probabilities, the remaining attributes of the complete development set are quantized into ten levels that are spaced uniformly. The pruned BN using individual features are shown in Figures 1. In Figure 2, we show the BN obtained by merging the pruned BNs obtained using individual features.

Table 1: MediaEval 2015 results on test set

	Affective impact (accuracies in %)		Violence detection (MAP)
	valence impact	arousal impact	violence
*run*1	35.66	44.99	0.0638
*run*2	34.27	43.88	0.0638
*run*3	33.89	45.29	0.0459
*run*4	29.79	48.95	0.0419
*run*5	33.37	43.97	0.0553

The configurations of five submitted runs (*run*1-*run*5) are the same for the two different subtasks. The first two *run* submissions (*run*1 and *run*2) are based on BN, third and fourth (*run*3 and *run*4) are based on ANN. The *run*5 results are obtained using random guess, based on the distribution of the samples in the development set. In *run*1, we have created a BN with all features (image, video and audio) by merging the networks learned individually using image features, video features and audio features respectively on the complete development data. In *run*2, a BN is created without audio features by merging the networks learned individually using image and video features on the complete development data. In *run*3 for the violence detection subtask, 19 different ANNs with openSMILE paralinguistic audio features (13 dimensional after feature selection) are trained. In *run*4 for the violent subtask, we have trained 19 different ANNs with 5 different set of features (41 dimensional MediaEval features, 20 dimensional audio MediaEval features, openSMILE audio features (7 dimensional after feature selection), openSMILE paralinguistic audio features (13 dimensional), and combination of openSMILE audio and MediaEval video and image features). So, we have trained $19*5 = 95$ ANNs. The best five ANN classifiers are selected while working on the development set. The development set is partitioned into 80% and 20% for training and testing, respectively. For the affective impact task, *run*3 and *run*4, we have trained several ANNs, each with a different feature set.

Table 1 shows the results with the metric proposed in MediaEval 2015 [6]. The best result (i.e. 48.95% accuracy) of affective impact detection is obtained with *run*4 for arousal detection that combines the best five neural networks for five different feature sets. And the best result (0.0638 of MAP) for violence detection is obtained in *run*2 that uses a BN.

3. REFERENCES

[1] A. Shah and P. Woolf, “Python environment for Bayesian Learning: Inferring the structure of Bayesian Networks from knowledge and data,” *Journal of Machine Learning Research*, vol. 10, pp. 159–162, June 2009.

[2] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[3] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning.* The MIT Press, 2009.

[4] “opensmile,” 2015. [Online]. Available: http://www.audeering.com/research/opensmile

[5] B. W. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *INTERSPEECH*, 2009, pp. 312–315.

[6] M. Sjoberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandrea, M. Schedl, C.-H. Demarty, and L. Chen, “The mediaeval 2015 affective impact of movies task,” in *MediaEval 2015 Workshop*, 2015.

[7] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, “Liris-accede: A video database for affective content analysis,” *IEEE Transaction on Affective Computing*, vol. 6, pp. 43–55, January 2015.