Brno University of Technology at MediaEval 2011 Genre Tagging Task

Michal Hradiš, Ivo Řezníček, Kamil Behúň
Graph@FIT, Brno University of Technology, Bozetechova 2, Brno, CZ
{ihradis, ireznice}@fit.vutbr.cz, xbehun03@stud.fit.vutbr.cz

ABSTRACT

This paper briefly describes our approach to the video genre tagging task which was a part of MediaEval 2011. We focused mainly on visual and audio information, and we exploited metadata and automatic speech transcripts only in a very basic way. Our approach relied on classification and on classifier fusion to combine different sources of information. We did not use any additional training data except the very small exemplary set provided by MediaEval (only 246 videos). The best performance was achieved by metadata alone. Combination with the other sources of information did not improve results in the submitted runs. This was achieved later by choosing more suitable weights in fusion. Excluding the metadata, audio and video gave better results than speech transcripts. Using classifiers for 345 semantic classes from TRECVID 2011 semantic indexing (SIN) task to project the data worked better than classifying directly from video and audio features.

1. INTRODUCTION

Our approach was mainly motivated by a question how video classification approaches which we employ to solve TRECVID SIN task [2] behave in a different context. The genre tagging task [3] is similar to SIN except the classes are of different kind, videos belong as a whole to a single class and, most importantly, the provided training set is in this case more than magnitude smaller.

We attempted to exploit most of the modalities available: video, audio, automatic speech recognition [1] (ASR) and user-supplied metadata. We did not use social network information. The image features extracted from video were standard Bag of visual Words (BOW) representations commonly used for image classification [5]. Spectrograms from audio were processed in the same way as image data. BOW was constructed from metadata and ASR as well.

2. METHOD

The BOW representation of video frames was constructed in a standard way [5] starting with local patch sampling followed by computing descriptors [4] and a codebook transform. We used Harris-Laplace detector (**HARLAP**) and dense sampling with position step 8 pixels and patch radius 8 pixels (**DENSE8**), respective 16 pixels (**DENSE16**). The extracted patches were represented by SIFT and color SIFT descriptors (**SIFT**, **CSIFT**). Codebooks were created for each representation by *k-means* with Euclidean distance and exact nearest neighbor search. The size of all codebooks was 4096. Local features were translated to BOW using codebook uncertainty [6] with Gaussian kernel and the standard deviation set to average distance between closest neighboring codewords. The BOW vectors were normalized to L_1 unit length for classification.

BOW representation from audio was extracted almost in the same way as from video. In this case the one-dimensional audio signal was converted to mel-frequency spectrogram which is a 2D representation and can be treated as an image. For spectrograms, only **DENSE8** and **DENSE16** sampling was used because the spectrograms do not contain distinct interest regions which Harris-Laplace could detect. Only **SIFT** descriptor was used as spectrograms do not contain any color information.

As the provided training set is extremely small, we decided to expand this set by treating each video frame and short spectrogram as individual sample with label equal to label of the original video, and merge these partial decisions later. 100 equidistant samples were extracted from each video (training and testing). The length of the spectrograms was set to 10 second and an overlap was allowed when needed. Linear SVM was used to learn separate 1-to-all classifiers for each genre. Meta-parameter C was set in cross-validation which asserted that samples from a single video did not appear in training and testing set at the same time. Considering the small number of original videos, we set the same C for all classifiers for a particular representation. The final response for a video was computed as the number of samples from that video for which the classifier for a particular genre gave the highest response compared to the other genre classifiers.

For TRECVID 2011 SIN task, we created classifiers for 345 semantic classes. These classifiers were created in almost the same way as the classifiers described in the previous text. We applied these 345 classifiers to the image and audio samples and created feature representations for the videos by computing histograms of their responses (8 bins per semantic class). The response histograms were then used to train genre classifiers as before - with a difference that the training sets were only the 246 videos and that the responses were directly used as results. These classifiers are further denoted with **TV11**.

For metadata and ASR we computed BOW representation by removing XML elements, non-alphabetic characters and

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

by splitting words where lower-case character was followed by upper-case character. Classifiers for metedata and ASR were created in the same way as for **TV11** representation.

We assumed that the number of training samples available is too small for accurate and reliable fusion. For this reason we decided to make an educated guess based on previous experience and results on the training set and combine classifiers by weighted average with the weights set by hand. Responses of classifier based on all audio and video features were averaged into **C-AV**, **TV11** into **C-TV11** and ASR and metadata into **C-TEXT**. These averages were normalized the biggest standard deviation of individual class responses and were combined by weighted average.

RUN1 used only ASR as required. **RUN3** combined **C-AV**, **C-TV11** and **C-TEXT** with weight for **C-TEXT** increased to 2.5. **RUN4** combined **C-AV** and **C-TEXT** which had weight 1.25. **RUN4** combined **C-TV11** and **C-TEXT** which had weight 1.25.

3. RESULTS

The results of the official runs are shown in Table 1. Using the MediaEval methodology, we additionally evaluated all the separate parts which were combined for the official runs as well as some other combinations. These unofficial results are shown in Table 2.

From the individual types of features, the best results were achieved by metadata. Metadata gives better results than all the official runs where adding other features decreased the performance. **TV11** classifiers provide significantly better results than classifiers trained directly on image features. The same is true also for their combinations, where **C-TV11** gives 0.275 MAP and **C-AV** gives only 0.226 MAP. Question remains if this is because the TRECVID classifiers bring additional knowledge or due to the differences in the training of the two sets of classifiers. Interestingly, the audio features provide good results comparable to visual features in **TV11**, and are much better than image features when learning directly from the features. The worse results in the case of **TV11** could be explained by lower performance of the original audio-classifiers on TRECVID data (almost two times worse than image features).

Further, we experimented with additional combinations of features. We combined all classifiers and all classifiers excluding METADATA with weights which more reflect performance of the classifiers. These result are denoted as **ALL**, respectively **ALL_WITHOUT_METADATA**, in Table 2. The weights were 1× **ASR**, 1× **C-AV**, 4× **C-TV11** and 8× **METADATA**. The combination **ALL** provides overall best result 0.448 MAP and significantly improves over the metadata alone. Improving over all its components, **ALL_WITHOUT_METADATA** reaches 0.3 MAP.

4. CONCLUSION

The achieved results are surprisingly good considering the small size of the training set used. Question is how the results would compare to other methods on this dataset, especially to those which use external sources of knowledge and which focus more on the metadata, as it was shown to be the most important source of information. Additionally, it is not certain how the presented methods would work on more diverse dataset.

Although, the metadata is definitely the most important source of information for genre recognition, the audio and video content features improved results when appropriately combined. A larger training set would be needed to perform proper classifier fusion which could further increase the benefit of the content-based features.

Run	MAP
RUN1	0.165
RUN3	0.346
RUN4	0.322
RUN5	0.360

Table 1: Mean average precision on test set achieved by the runs submitted to MediaEval 2011.

Features		TV11
DENSE16_CSIFT	0.126	0.194
DENSE16_SIFT	0.100	0.178
DENSE8_CSIFT	0.116	0.201
DENSE8_SIFT	0.078	0.187
HARLAP_CSIFT	0.145	0.178
HARLAP_SIFT	0.133	0.174
SPECTRUM_DENSE16_SIFT	0.195	0.167
SPECTRUM_DENSE16_SIFT	0.158	0.188
COMBINED (C-AV, C-TV11)	0.226	0.275
ASR	0.165	
METADATA	0.405	
C-TEXT	0.300	
ALL_WITHOUT_METADATA	0.300	
ALL_WITHOUT_ASR	0.448	
RANDOM	0.046	

Table 2: Unofficial results on testing set. Mean average precision reported.

Acknowledgements

This work has been supported by the EU FP7 project TA2: Together Anywhere, Together Anytime ICT-2007-214793, grant no 214793.

5. REFERENCES

[1] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2):89 – 108, 2002.

[2] Michal Hradis et al. Brno university of technology at trecvid 2010. In *TRECVID 2010: Participant Notebook Papers and Slides*, page 11. National Institute of Standards and Technology, 2010.

[3] Martha Larson et al. Overview of mediaeval 2011 rich speech retrieval task and genre tagging task. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

[4] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[5] Cees G. M. Snoek et al. The mediamill trecvid 2010 semantic video search engine. In *TRECVID 2010: Participant Notebook Papers and Slides*, 2010.

[6] J. C. van Gemert et al. Visual word ambiguity. *PAMI*, 32(7):1271–1283, 2010.