TUD at MediaEval 2012 genre tagging task: Multi-modality video categorization with one-vs-all classifiers

Peng Xu[1], Yangyang Shi[2], Martha A. Larson[1]
[1]*Delft Multimedia Information Retrieval Lab,* [2]*Interactive Intelligence Lap*
Delft University of Technology
Mekelweg 4, Delft, The Nethelands
{*p.xu@tudelft.nl, yangyangshi@ieee.org, m.a.larson@tudelft.nl*}

ABSTRACT

In this paper, we investigate the internet video categorization problems on genre related labels. The videos are represented by features extracted from different modalities. Then for each category, one-vs-all SVM classifiers are trained based on features from different modalities. The weighted Reciprocal Rank Fusion method is used to combine the classifiers for each modality. The experiments are based on the data from the Genre Tagging Task of MediaEval 2012.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Algorithms, Performance, Experimentation.

Keywords

Genre tagging,video classification, video representation

1. INTRODUCTION

In the Genre Tagging Task of MediaEval 2012 [4], the genre related tags are required to be predicted for a collection of internet videos. The challenge of this task lies in two aspects, the high diversity of the content for internet videos as well as the high semantic level of the genre related tags. This paper is to investigate the discriminative power of information from different modalities, such as visual features, speech transcriptions [3] and metadata (e.g., titles, descriptions, tags and uploader IDs). The focus is on analyzing the performance of visual feature for categories with different properties. Moreover, the combination of different modalities is investigated by a late fusion method.

2. APPROACHES

For each category, the one-vs-all SVM classifiers are trained based on the information from different modalities, including visual features, Automatic Speech Recognition transcripts and metadata. The classification results are ranked by their confidence scores. The ranked lists from different modalities are fused by a weighted Reciprocal Rank Fusion method. A post-processing procedure is applied to guarantee only one category can be assigned to each video.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

2.1 Visual feature

Vector quantized keypoint descriptors are used for representing visual information. The SIFT descriptors are extracted from each key-frame with densely sampled keypoints. Since the automatic key-frame extraction is not perfect, we regularly sampled 5 additional frames from the videos which have less than 5 key-frames. The hierarchical k-means clustering algorithm is used to cluster the descriptors from the development set into a visual vocabulary. In this algorithm, the clusters are organized into a tree structure and the exact number is difficult to determine. We set the algorithm terminates when there are at least 2000 clusters, which result in a vocabulary with 2187 bins. Then each video is represented by the term frequency of all the visual words and normalized by the number of key-frames in the video.

We also extract global video features: edit features consisting of mean and variance of shot length(2D), ratio of hard cut(1D); content features, which are constructed by first extracting 5(D) color features (the average brightness, saturation end entropy of the frame, the pixels whose brightness and saturation are higher than a predefined thresholds) and 5(D) GLCM based texture features from all the key-frames, and then representing the entire video by the mean and 2-order moments of each feature. This representation performs worse than the visual words based representation, so we did not report it in the official runs. Instead, we use it in an additional run for comparison.

2.2 Text feature

We base our approach on the assumption that speech information such as lexical items and syntactic structure are related to genre. We exploit the 1-best hypotheses from automatic speech recognition [3]. Latent Dirichlet Allocation (LDA)[1] was applied to a version of the 1-best hypotheses post-processed to retain semantically salient words which was obtained by removing the stopwords and punctuation from the original 1-best hypotheses. In this way, each transcript was mapped to a low dimensional latent topic vector. The weight is determined by the conditional probability of latent topics given the transcript.

The metadata is processed in a similar fashion. First, the video is represented by the vocabulary with the stemmed words from the development set. Second, the LDA is performed to reduce the feature dimensions. From the experience of last year [5], it is expected that exploiting all the information from title, descriptions and tags performs best.

In both the ASR and the metadata, the number of latent topics is estimated using the development set.

2.3 Classification and Fusion

For each category, the videos from the category are considered as the positive examples, while all the other videos are used as negative ones. A one-vs-all SVM classifier with linear kernel is trained based on features from every modality. The soft margin parameter C is determined by the 5-folder cross validation with in the development set. The ranked list is achieved by the confidence scores.

The ranked lists of different modalities are fused by the weighted Reciprocal Rank Fusion (RRF) [2]. Given a set D of documents to be ranked and a set of rankings R, the RRF scores are

$$RRFscore(d \in D) = \sum_{r \in R} w(r) \frac{1}{k + r(d)} \quad (1)$$

where $r(d)$ is the ranking score for each document in the rank R. k is a parameter to balance the importance of the higher ranked samples as well as the lower ranked samples. We choose $k = 60$, following the suggestion in [2]. The weight $w(r)$ is determined by the Average Precision of each classifier in the cross validation process.

In order to fulfill the requirement that only one category can be assigned to each video, a post-processing procedure is perform when the ranked list of all the categories are generated. In this step, the category label of certain video is assigned by choosing the category for which it has the highest ranking score.

3. RESULTS

The five official runs for this task are organized as followed: *run_1* the visual only results; *run_2* the ASR transcripts based results; *run_3* the visual feature combined with ASR transcripts; *run_4* the combination of visual features, ASR transcripts and the metadata; *run_5* based on the ranking list of *run_4*, the uploader IDs are used to increase the confidence scores for the videos in the test set associated with the same uploader as the videos in the development set for a given genre label.

In the unofficial *run_6*, we use the visual only features the same as run_1. The difference lies in that the ranked list for each category includes all the videos in the test set. In *run_7*, the global based visual feature is used, and the post-processing is performed.

The Mean Average Precision (MAP) is shown in Table 1.

Table 1: MAP of different runs (The unofficial runs are marked by *)

	run_1	run_2	run_3	run_4	run_5	run_6*	run_7*
MAP	0.0061	0.3127	0.2279	0.3675	0.2157	0.0577	0.0047

From Table 1, it can be seen that the text based information yields better performance than the visual based features. Since the low MAP of the visual only features (*run_1*), the combination of ASR and visual features (*run_3*) do not achieve any improvement with respect to the (*run_2*). The best performance of these 5 runs is the *run_4* which explores the combination of visual features, ASR and metadata. This result is consistent with last year's findings, which demonstrated the discriminative ability of metadata [5]. Meanwhile, in the late fusion stage, the weak classifiers of the visual features are weighted much lower in *run_4*. Given the results of last year, it is unexpected to see that the usage of uploader IDs does not improve the MAP in *run_5*. We conjecture that the lack of improvement is due to the weights for the uploader IDs were assigned without cross-validation. This results in the over high impact of uploader IDs, and the other modalities are neglected.

For the visual only result *run_1*, although the MAP is higher than the random baseline ($MAP = 0.0022$), it is still much lower than the classifiers from other modalities. There are mainly two reasons. First, the information from the visual channel is not discriminative enough to predict the genre related tags, which have high-level semantic interpretations. Second, to fulfill the requirement that only one category can be sent to each video, a post-processing procedure is perform based on the results of the one-vs-all classifiers. This procedure is far from optimum. One of the problems is that the prior probabilities of categories are not taken into account. Considering the fact that the dataset for this task is highly unbalanced, it is possible that the categories with small size are assigned too many samples. We can assume that for this reason the performance of other categories is deteriorated. It shows that on 20 out of 26 categories in *run_6*, the APs are higher than the random baseline, while only 15 categories are better than random in *run_1*.

In this dataset, videos with the same uploader ID tend to belong to the same series. Videos from the same series share certain visual similarities. So if most of the true positive videos for a certain query are from one or several shows, the categorization with visual features can achieve reasonable results, such as the query *1014 literature* and *1018 politics*. Since videos in the same series tend to have visually similar parts, which can be well measured by the visual words based features, the visual words based feature performs better in categories of which videos are mainly from several series(*1013 health* and *1014 literature*), while worse in categories from which series are more diverse (*1025 travel*).

4. CONCLUSIONS

In contrast to last year, the larger development set make it possible to train classifiers with visual feature only. So that the comparison between visual features and text features can be made directly. Although the experimental results for that the visual information is satisfactory, it is possible to improve in several ways. First, the visual words based representation lies a high dimensional space with respect to the relatively small training set. Therefore, a appropriate feature selection or dimensional reduction method may improve the results. Second, the post-processing procedure to assign one category label for each video should be improved by taking into account of the prior of each category. Third, it is still worth exploring other video representations which can be more discriminative to the genre level categories.

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. SIGIR '09, pages 758–759. ACM, 2009.

[3] A. Rousseau, F. Bougares, P. Deleglise, H. Schwenk, and Y. Esteve. Lium systems for the iwslt 2011 speech translation tasks. In *International Workshop on Spoken Language Translation*, 8-9 Sept 2011.

[4] S. Schmiedeke, C. Kofler, and I. Ferrané. Overview of MediaEval 2012 Genre Tagging Task. In *MediaEval 2012 Workshop*, Pisa, Italy, Octobor 3-4 2012.

[5] P. Xu, D. Tax, and A. Hanjalic. TUD-MM at MediaEval 2011 Genre Tagging Task: Video search reranking for genre tagging. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.