TUB @ MediaEval 2011 Genre Tagging Task: Prediction using Bag-of-(visual)-Words Approaches

Sebastian Schmiedeke, Pascal Kelm and Thomas Sikora
Communication Systems Group
Technische Universität Berlin, Germany
{schmiedeke,kelm,sikora}@nue.tu-berlin.de

ABSTRACT

This paper describes our participation in Genre Tagging Task of MediaEval 2011, which aims to predict the videos' category label. We use bag-of-words approaches with different features derived from visual content and associated textual information. We perform different experiments in which different constellations in respect of single modalities, classification methods, visual features and their combinations are investigated. Each video of the test set is assigned to a single genre label, therefore, the classification accuray (CA) is a good metric for evaluation. As expected the most pieces of information for distinguishing genre contain the metadata (MAP = 0.2988 / CA = 65%). In combination with visual words the performance can be increased (MAP = 0.3033 / CA = 65.2%).

Categories and Subject Descriptors

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

General Terms

genre recognition, media annotation, bag of words

1. INTRODUCTION

The possibilities arising from new technologies (such as Web 2.0) facilitate significantly the production and dissemination of new content. Automatic classification of video is needed due to the huge amount of data and to enable user to easier find the desired content. Recent approaches are surveyed in [1]. The Genre Tagging Task as part of MediaEval 2011 required participants to automatically predict genre labels for a set of user generated Internet videos from blip.tv. These video sequences are accompanied by automatic speech recognition (ASR) generated transcripts provided by LIMSI-CNRS, various metadata–such as titles, descriptions, comments, tags–and key frames from shots. The data set is described in [2].

2. FRAMEWORK & METHODOLGY

Our proposed framework includes textual and visual information of shared media. Due to varying character of the videos the metadata and transcripts are available in different languages. For this reason we detect the language and we translate the text into English using the web service Google Translate[1]. These words are stemmed with an implementation of Porter's algorithm[2]. Next step is the removal of stemmed English stop words. For representation we choose the Bag-of-Words (BoW) model and apply classifiers that work well with the this representation. The textual vocabulary V is built with stemmed words of the development set filtered for stop words. The words of the test set are mapped to this vocabulary; words appearing only in the test set are discarded. Then, a term vector $t(\vec{d})$ is generated for each video that indicates its frequency of terms t. All videos from the development set form the term-document matrix $D_{dev}(d,t)$ These term vectors are classified with the following methods:

(1) Multi-class support vector machine (SVM) with linear kernel and cost parameter $C = 1$. The classification into multiple genres is obtained using the "one-vs.-one" strategy and the majority voting rule.

(2) Multinomial Naive Bayes (NB) with add-one smoothing; the core is the probability $P(t|c)$ that contains the probability of each term occurrence per genre:

$$P(t|c) = \frac{1 + \sum_{\forall d' \in c} D_{dev}(d', t)}{\sum_{\forall d' \in c} \left[1 + \sum_{\forall t' \in V} D_{dev}(d', t')\right]}$$

The decision to a genre for a video is obtained by:

$$c_v = argmax_{c \in \mathbb{C}} \left[\sum_{i=1}^{|V|} D_{dev}(d_v, t_i) P(t_i|c) \right],$$

In order to avoid floating point underflow, the multiplications of these probabilities are replaced with additions of their logarithms. The a-priori knowledge $P(c)$ is ignored assuming the genre distribution does not match within the sets.

(3) Nearest-neighbour classification (NN) of Jensen- Shannon divergence on term vectors; every term vector of the test set is compared to each term vector of the development set using this distance:

$$dist(i,j) = JSD(t(\vec{d}_i), t(\vec{d}_j)).$$

Each test video gets the genre label of the training video with the smallest distance. In case of SVM classification and this nearest-neighbour classification, the term vectors are weighted with term frequency and inverse document frequency (Tf-idf).

[1] http://translate.google.com
[2] http://tartarus.org/~martin/PorterStemmer

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

Table 1: Results on official submitted run; MAP and CA

experiment	input	classification	MAP	CA
run1	translated ASR	NN + Jensen-Shannon-Divergence	0.1824	48.41%
run2	Metadata without tags	Naive Bayes	0.2986	65.04%
run3	Metadata, translated ASR	Naive Bayes + serial fusion	0.3049	66.64%
run4	Metadata, clustered SURF	Naive Bayes, linear SVM + serial fusion	0.3033	65.22%
run5	clustered SURF	SVM with RBF kernel	0.0943	43.68%

The visual content is described by local features (SURF) extracted from each key frame of video sequences. Each key frame is described by several SURF features at key points and at a regular grid. These entire local features extracted from gray-scale versions of the key frames of the development set are clustered to get a 2048-sized vocabulary. This vocabulary is the basis for generating the term vectors for each key frame. A representation for a single video is obtained by bin wise pooling of the key frames' term vectors. The resulting term-document matrices are weighted with Tf-idf. Then, these bags of visual words are classified with a SVM.

3. EXPERIMENTS & RESULTS

We perform several experiments which differs in the use of ressources and classification methods. The results of the official runs are depicted in table 1.

3.1 Textual features

Run1 and **run2** are performed on ASR transcripts and metadata (without tags) only with different classification methods used. The effect of these classification methods applied on non-translated metadata is shown in table 2. SVM

Table 2: Evaluation of classification methods

	SVM (linear $C = 1$)	Naives Bayes	JSD+NN
MAP	0.1874	0.2989	0.2585
CA	54.31%	65.04%	62.84%

performs worse than Naive Bayes, this is maybe due to the sparseness of the term vectors. A SVM with RBF kernel and proper parameter may achieve better results. **Run3** combines non-translated metadata and translated transcripts in a serial way by first relying on metadata and when falling below a certain confidence score using ASR transcripts. This fusion works better than the single ressources or their combination, as shown in table 3. This table shows that metadata

Table 3: Evaluation of (translated) textual resources; Naive Bayes

	ASR	metadata	combined
MAP	0.0703	0.2853	0.0783
CA	48.41%	63.88%	49.68%

contains more discriminative power than using additional speech transcriptions.

3.2 Visual features

Run5 is a purely visual approach in which a SVM with RBF kernel with parameter found by cross validation parameter search is used. The term vectors of the 5th run are pooled by averaging. The Evaluation of pooling methods is shown in table 4. The average pooling of single key frame's

Table 4: Evaluation of pooling methods; SVM with linear kernel $C = 1$

	max	avg	median	no/fusion
MAP	0.0673	0.0845	0.0003	0.0686
CA	38.4%	40.84%	1.16%	39.68%

term vector achieves the best result, also in comparison to the fusion of decision on single key frames.

3.3 Fusion

BoW representations of textual and visual features can be easily combined by concatenating the respective term vectors. An equal-ranking of these features decrease the performance compared to textual feature alone, as shown in table 5. A serial fusion like in **run4** achieves better results by

Table 5: Evaluation of direct fusion

	visual feat. only	text feat. only	feat.level fusion
MAP	0.0673	0.1839	0.1203
CA	38.4%	53.91%	46.61%

first relying on metadata and when falling below a certain confidence score using visual features.

4. CONCLUSION

We demonstrate a simple, but efficient approach to predict genres which is based on supervised classification methods. We show that metadata contains the most discriminative power for distinguishing these genres. In a sophisticated fusion, visual feature can make a contribution to better results.

5. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's FP7 under grant agreement number 216444 (NoE PetaMedia).

6. REFERENCES

[1] D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):416 –430, may 2008.

[2] M. Larson, M. Eskevich, R. Ordelman, C. Kofler, S. Schmiedeke, and G. Jones. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.