TUB @ MediaEval 2012 Tagging Task: Feature Selection Methods for 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 the Genre Tagging Task of MediaEval 2012, which aims to predict the videos' category label. In last year's participation, we performed experiments with bag-of-words (BoW) approaches in which different constellations in respect of modalities, features, and methods were investigated. This year, we focus on feature selection methods to improve the classification performance in terms of mean average precision (mAP) and classification accuracy (CA). We investigated the effectiveness of selection methods based on scores calculated using mutual information (MI) or term frequency (TF) and the effectiveness of transformation methods like the principle component analysis (PCA).

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Experimentation

Keywords

genre classification, bag-of-words, SURF, Feature selection

1. INTRODUCTION

We address the issue of automatic genre labelling of web videos, since manual annotation is laborious due to the huge amount of newly generated data. The data set contains among the visual content, shot boundary information [1], automatic speech recognition (ASR) transcripts [2] as well as social and textual metadata. The whole set with its 26 genres is described in the task overview paper [4]. An overview of state-of-the-art literature can be found in [3]. This working notes paper is structured as follows: We introduce our approach using feature selection (FS) methods applied on BoW features. The results are then shown in section 3, followed by a conclusion summarizing our main findings.

2. METHODOLOGY

Our proposed framework includes textual and visual information of shared media. The characteristics of the data set is presented using the BoW approach. Therefore, words from the metadata and ASR transcripts are stemmed using the Porter stemmer algorithm [1], once stop words and digits were removed. These stemmed words build the textual vocabularies V_{TXT} and V_{ASR}, by analogy the visual vocabulary V_{VIS} is built from clustered local features (SURF). These SURF are extracted from densely-sampled and sparsely-sampled keypoints which are then clustered hierarchically to get the 4096-sized (8092-sized respectively) vocabulary. In contrast to the textual vocabularies, in V_{VIS} each terms vector represents a (key) frame instead of the whole sequence. So, a representation for a whole video is obtained by bin-wise pooling of each key frames' term vectors. For more details see [3].

2.1 Feature Selection/ Transformation

The terms are filtered for that one with the most discriminative power. The discriminative power of each term is indicated as the measure $A(t,c)$ in a two-class problem. The higher the score of $A(t,c)$, the more important is the term t for class c. The measure can be determined in multiple ways:

Mutual information (MI) measures how much information the presence(1)/ absence(0) of a term t contributes to the correct decision on class c. MI is defined by:

$$A(t,c) = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{N_{i,j}}{N} \log_2 \frac{N \cdot N_{i,j}}{N_i \cdot N_j},$$

where N denotes the count of all video sequences. The subscript i indicates the presence of term t and j indicates the membership to class c (e. g. $N_{1,0}$ is the count of videos that contains the term, but does not belong the class c). Further, $N_i = \sum_{j=0}^{1} N_{i,j}$ is the number of videos containing term t and $N_j = \sum_{i=0}^{1} N_{i,j}$ is the count of videos belonging to class c. The mutual information reaches a maximum value, when the relative term is only found in a single class.

Term frequency (TF) is a quite simple measure and it is defined by the number of the term t occurring within a certain class c.

Among these selection methods, we also investigate the performance of transformation methods like the *principle component analysis* (PCA). The PCA performs an eigendecomposition of the co-variance matrix $\mathbf{C}$ of the term vectors $\mathbf{X}$. As with the scores of the selection methods, the higher the score of eigenvalue λ_i, the more is the effect of eigenvector v_i on the total variance. So, the eigenvectors are

[1] http://tartarus.org/~martin/PorterStemmer

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

Table 1: Results on official runs—evaluated using mAP and CA

Run	Name	Feature	FS	Classification	mAP	CA
1	SURF_PCA	clustered SURF $\|V_{VIS}\| = 8192$	PCA	SVM (HI)	0.2301	41.63 %
2	ASR_FS:TF	V_{ASR} (LIMSI)	TF	Naive Bayes	0.1035	32.53 %
3	SURF_FS:MI	clustered SURF $\|V_{VIS}\| = 4096$	MI	SVM (HI)	0.2259	40.80 %
4	TEXT_FS:MI	tags, title, description (V_{TXT})	MI	Naive Bayes	0.5225	58.18 %
5	SURF_PCA,SUSC	like in 1 + uploader name	PCA	SVM (HI) + CV	0.3304	52.14 %

sorted in descending order by their eigenvalues which denote the transformation matrix $\mathbf{W}$, respectively shorten by low-valued eigenvectors. The feature space is transformed by matrix multiplication: $\mathbf{X_{PCA}} = \mathbf{W}^{\mathrm{T}} \cdot \mathbf{X}$.

The selection for the most important term is based on different approaches:

(Top-k_Union): Select the union of the top k terms with corresponding score sorted in descending order per class.

(Top-k): Select the top k terms pooled (max, min, avg) over all classes.

(Union> Th): Select the union of all terms where its values $A(t, c)$ exceeds a threshold th (e. g. , mean value) in any class.

(Intersection> Th): Select the intersection of all terms where its values $A(t, c)$ exceeds a threshold th for all classes.

2.2 Classification

These reduced/ transformed term vectors are then classified with the following methods:

(1) Multi-class support vector machine (SVM) with histogram intersection (HI) kernel and cost parameter $C = 1$. The classification into multiple genres is obtained using the *one-vs-one* strategy and the majority voting rule. The HI kernel is defined by $\kappa(\vec{x}, \vec{y}) = \sum_i \min(x_i, y_i)$.

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

$$P\left(t|c\right) = \left(N_{t,c} + 1\right) / \sum_{t' \in V} \left(N_{t',c} + 1\right),$$

where $N_{t,c}$ is the term occurrence of term t in class c. Smoothing is necessary to have a probability value higher than zero for all terms in all classes. For each video the class probabilities are calculated by addition the logarithms of $P(t'|c)$. The logarithms are used to avoid floating point underflow. The decision is then obtained by choosing the class with the highest probability. Since the scores are important for mAP calculation, the logarithmic class probabilities are exponentiated and then normalised ($\sum = 1$).

3. EXPERIMENTS & CONCLUSION

We perform the following official runs, as shown in table 1:

(1) **SURF_PCA**: The results of this run are classified using a SVM with histogram intersection kernel. The feature is here a bag of SURF that has been transformed using PCA. Here, the feature dimensionality is not reduced.

(2) **ASR_FS:TF**: In this run a bag of filtered words coming from LIMSI's speech transcripts (V_{ASR}) is applied in a Naive Bayes classifier. Words are filtered using term frequency and *Union> Th* method.

(3) **SURF_FS:MI** uses a bag of filtered SURF ($|V_{VIS}| = 4096$) as feature. Only those ”visual” words are selected which mutual information of every class exceed the overall mean value (*Intersection> Th*).

(4) **TEXT_FS:MI**: Here a bag of words from the metadata (V_{TXT}) are filtered for words which mutual information exceed the overall mean value in any class (*Union> Th*). Then, this feature is applied to the Naive Bayes classifier.

(5) **SURF_PCA,SUSC**: The same as the first run (SURF _PCA), but uploader information (prefix of file name) is used to get a single genre decision (consensus voting (CV)) for videos of the same uploader.

As expected, the best result is achieved using filtered metadata as described in run 4 (mAP=0.5225); the mAP score is increased by 0.1119 campared to an run using all metadata. The best run, which uses only visual content achieves a mAP of 0.2301 and can be improved by using additional uploader information up to a mAP of 0.3304. The performance differences applying selection methods is less significant for the visual runs, likely caused by the fact that the visual vocabulary is much smaller than the textual ones. So, the first run does not benefit from the feature transformation—the mAP is slightly decreased by 0.001—while run 5 is the visual run which benefits most from the transformation. Here, the mAP increases by 0.031 compared to an unofficial run using the same configuration, but not applying PCA.

As the results showed feature selection methods are able to improve results, although the determination of the threshold parameter is essential. The choice of using MI or TF is not critical, both methods achieve roughly the same results, but TF is much faster to compute. In future we investigate different scaling schemes applied to our Bayesian model.

4. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's FP7 under grant agreement number FP7-261743 (VideoSense).

5. REFERENCES

[1] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In *10th Workshop on Image Analysis for Multimedia Interactive Services.*, 2009.

[2] L. Lamel and J.-L. Gauvain. Speech processing for audio indexing. In *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 4–15. Springer Berlin Heidelberg, 2008.

[3] S. Schmiedeke, P. Kelm, and T. Sikora. Cross-modal categorisation of user-generated video sequences. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pages 25:1–25:8, 2012.

[4] S. Schmiedeke, C. Kofler, and I. Ferrané. Overview of MediaEval 2012 Genre Tagging Task. In *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012.