TUD-MM at MediaEval 2011 Genre Tagging Task: Video Search Reranking for Genre Tagging

Peng Xu[1], D.M.J. Tax[2], Alan Hanjalic[1]
[1]Delft Information Retrieval Lab, [2]Pattern Recognition Laboratory
Delft University of Technology
Mekelweg 4, Delft, The Nethelands
{p.xu, d.tax, a.hanjalic@tudelft.nl}

ABSTRACT
In this paper, we investigate the possibility of using visual information to improve the text based ranking. Both a structure based representation (using the similarity matrix of the frames of one video) as well as a key-frame based representation (using visual words) is evaluated. It appears that only in some queries the visual information can improve the performance by reranking. The presented experiments on reranking show the limitations and also the potential, for these structural and visual representations

Categories and Subject Descriptors
H.3 [**Information storage and Retrieval**]: H3.1 Content Analysis and Indexing; H3.7 Digital libraries.

General Terms
Measurement, Performance, Experimentation,

Keywords
Video representation, Video search reranking

1. INTRODUCTION
In the MediaEval 2011 Genre Tagging Task, internet videos have to be ranked according to their relevance for a set of genre tags [1]. For certain domains, visual information has been proved to be related to video genres, but it is still a challenging problem for internet videos, because of the diversity of the video content.

This research evaluates the potential of visual information for genre level video retrieval. This problem is addressed by using visual information to rerank the text retrieval ranking list. Despite the different strategies used in various reranking methods, the basic assumption is that visually similar videos should be ranked in nearby positions in the ranking list. Therefore, it is important to find the appropriate visual features to represent movies.

2. VIDEO REPRESENTATIONS
In this paper, a Bayesian reranking approach is performed based on two kinds of video representations: the first one is a structure based feature and the second one is a key frame based feature.

2.1 Structure based representation
Most video measurements are based on comparing visual similarity between videos directly, using color, shapes and movement. However, these measurements would not always work due to the high variance of video content. In particular, video from the same genre are not expected to have the same visual content. Motivated by the limitations of visual similarity, we proposed a method for video structure representation and measurement.

In this method, a video self-similarity matrix is used to represent the structure. This matrix is generated by calculating the pairwise similarity between frames from one video, that are sampled with a fixed sampling rate. These similarities are calculated from the HSV histogram of each frame. This representation exploits the fact that one video tends to have consistent quality and editing conditions. A reliable similarity can be achieved without complicated low level visual features or additional domain knowledge

Each video is now represented by a square similarity matrix, which can be considered as a square gray-level image. Next, three types of multi-scale statistical image features can be extracted from the self-similarity matrix: a) GLCM based features (30 components). The Gray Level Co-occurrence Matrix (GLCM) is constructed from the similarity matrix for 2 directions (0, 45) and 3 offsets (3, 6, 12). For each GLCM, energy, entropy, correlation, contrast and Homogeneity are computed; b) 3-scale Gabor texture features (14 components). c) Intensity Coherence Vectors (16 components). Each pixel within a given intensity bin is classified as either coherent or in coherent, based on whether or not it is part of a large similarly-colored region. The size of the region is determined by a fix threshold. (1/15 of the size of image is used.)

2.2 Key-frame based representation
Next to the structure, a *Visual Word* representation based on key-frames is used for measuring the visual similarity between videos [2]. The key-frames are clustered into N clusters using K-means clustering based on image features. Next, every key-frame can be assigned to a cluster label, and the label histogram is finally used as the representation of the video. This feature was designed for web video categorization. The number of clusters is set as *K=400*. Additional experiments have shown that the performance is not sensitive to this parameter.

3. BAYESIAN RERANKING
The attractiveness of reranking is that it is naturally unsupervised. Given an initial ranking list, an improved one can be achieved by grouping the visually similar videos into the nearby positions. However, in practice, the designing of re-ranking methods and the setting of parameters are highly depended on the quality and characteristics of the base line ranking list.

Bayesian video re-ranking method is used in this paper, because it requires less assumptions of the original ranking list and it is less sensitive to particular parameter settings [3]. In this method, the

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

reranking problem is considered as minimizing the following energy function.

$$E(r) = \frac{1}{2}\sum_{i,j} w_{ij}(r_i - r_j)^2 + c\sum_{(i,j)\in S_T}(1 - \frac{r_1 - r_2}{\bar{r}_1 - \bar{r}_2})$$

Here $r = [r_1, r_2, \cdots, r_N]$ is the ranking list after reranking, $\bar{r} = [\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_N]$ is the initial ranking list, $w_{ij} = \exp(-\|x_i - x_j\|/\sigma)$ is the visual similarity between two items in the refined ranking list. The first term measures the visual consistency of the ranking list, while the second term refers to the ranking distance between the reranking list and initial list. *c* is a trade-off parameter to the two terms, which can be optimized on the development set.

4. RESULTS

Reranking is performed based on two baseline ranking lists which are generated by text retrieval. The first one uses information of automatic speech transcript (ASR), the second one uses metadata of the video. The second one is expected to outperform the first one. Details of generating these two baselines can be seen at [4].

Five official runs for this task are submitted. The 5 runs are organized as followed: 1) Gabor feature combined with CCV feature on ASR baseline; 2) Visual words based feature on ASR baseline; 3) Gabor feature combined with CCV feature on metadata baseline; 4) Visual words based feature on metadata baseline 5) GLCM feature combined with CCV feature on metadata baseline. The comparison of Mean average precision (MAP) of text baseline and reranking results are shown in Table.1.

Table 1. MAPs of text baseline and re-ranking results

Baseline		ASR	Metadata
		0.2146	0.3936
Reranking	Gabor+CCV	0.2060	0.3703
	GLCM+CCV	---	0.3690
	VW	0.2098	0.3605

It can be seen in Table1 that compared with the initial ranking lists, the reranking process did not improve the overall MAP. This result is unexpected, because there are some results in literature that suggests that the visual channel may contain information about the video genre [5]. It appears that in this dataset, around one fourth of the videos contain a single person talking with little visual aids. Therefore, these videos do not contain sufficient information in the visual channel to estimate the genre tags of these videos.

Furthermore, many videos in this dataset are presented in series. 1390 videos in the test set have more than 2 episodes belong to the same show. Videos of the same show tend to share certain visual similarities. Through the analysis of the reranking performance on each query, it can be observed that this property has a strong effect on the performance. (The reranking results for some selected queries are presented in Figure 1.) Generally speaking, if most of the true positive videos for a certain query are from one or several shows, the reranking results can be reasonable, such as the query '1016 politics'.

In particular, the most significant improvement appeared in the query '1001 autos_and_vehicles'. It can be seen from the ground truth that all the 6 videos in this genre are episodes of a same show. The reranking process takes advantage of the high visual similarity between the 6 videos. Especially, for this query, key-frame based features achieved higher performance than the structure based features. This is because videos in the same series tend to have duplicate parts, which can be easily detected by visual similarity based representations.

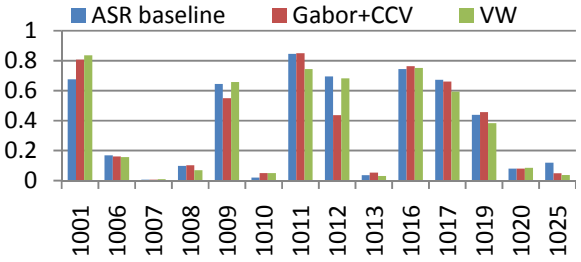

Figure 1. Reranking results for selected queries on ASR text baseline

Exploring the similarity within the same show is not enough for genre tagging. Videos of the same show have certain chances to be of different genres. Moreover, there are some queries of which videos are from many different shows. (For example, the 64 videos of the query '1019 sports' are from 16 different shows.) In these cases, the visual similarities between true positive videos are not obvious. Therefore, the structure based features outperform key-frame based ones. This indicates that the videos of same genre may share similarities in structure even though they are not consistent in visual.

5. DISCUSSION AND FUTURE WORK

Although the visual reranking made no improvement for the initial ranking list on MAP, it does not necessarily mean that visual information is useless for detecting video genre. It is the special characteristics of this dataset that make it difficult to utilize information in visual channel. In particular, compared with conventional understanding of video 'genres', the genre tags given in this task are more related to the 'topics' of videos. The proposed structure based video representation provided a possibility for an inexact matching for video similarity. The characteristics of this representation can be observed through analyzing the reranking performance of certain queries. It is still not clear what is the most suitable way of representing the frame similarity matrix. A more attractive direction may be discovering a set of tags which could reflect the visual consistency of videos.

6. REFERENCES

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

[2] Yang, L., Liu, J., Yang, X and Hua, X-S. 2007. Multi-modality web video categorization. In *Proceedings of MIR '07*. 265-274

[3] Tian, X., Yang, L., Wang, J., Yang, Y., Wu, X. and Hua, X.-S. 2008. Bayesian video search reranking. In *Proceeding of MM'08, ACM,* 131-140

[4] Rudinac, S., Larson, M., and Hanjalic A., 2011. TUD-MIR at MediaEval 2011 Genre Tagging Task: Query Expansion from a Limited Number of Labeled Videos, *In Working Notes MediaEval 2011*

[5] Brezeale, D. and Cook, D. Automatic Video Classification: A Survey of the Literature. 2008. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 416 -430