MediaEval 2013: Soundtrack Selection for Commercials Based on Content Correlation Modeling

Han Su[1], Fang-Fei Kuo[2], Chu-Hsiang Chiu[1], Yen-Ju Chou[1], Man-Kwan Shan[1]

Department of Computer Science, National Chengchi University, Taipei, Taiwan[1]

Department of Electrical Engineering, University of Washington, Washington, America[2]

{101753004,101753026,101971001,mkshan}@nccu.edu.tw[1], ffkuo@uw.edu[2]

ABSTRACT

This paper presents our approaches of soundtrack selection for commercials based on audio/visual correlation analysis. Two approaches are adopted. One is based on multimodal latent semantic analysis (MLSA) and the other is based on cross-modal factor analysis (CFA). The evaluation based on the MediaEval Soundtrack Selection for Commercials Dataset shows the performance of our systems.

Keywords

Soundtrack selection, Multimodal correlation analysis, Multi-type latent semantic analysis, Cross-modal factor analysis

1. MOTIVATION

Automatic soundtrack selection for videos has received more and more attention. The rationale of our approach for automatic soundtrack selection is based on the latent correlation of the video and audio from training data (Development Dataset). Two methods of multimodal correlation model learning are utilized in our approach. In this paper, we present our soundtrack recommendation using the two methods respectively and evaluate the system on the MediaEval corpus.

2. SYSTEM ARCHITECTURE

Figure 1 shows the architecture of the proposed soundtrack selection based on our previous work [1]. In the training phase, we first transform the descriptors of audio/visual features provided in the development dataset (*devset*) to the **audio /visual words** and generate the **audio/visual feature matrices**. Then two algorithms are employed to find the **content correlation model** from the visual/audio feature matrices. For the recommendation dataset (*recset*), the audio features of each soundtrack are transformed into audio words in the same way as the development dataset do. In the test phase, given a test video, the descriptors of visual features are transformed into visual words in the same way as those of the *devset* The transformed visual words of the test video along with the audio words of *recset* are fed into the learned content correlation model and the ranking results for soundtrack selection are generated.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

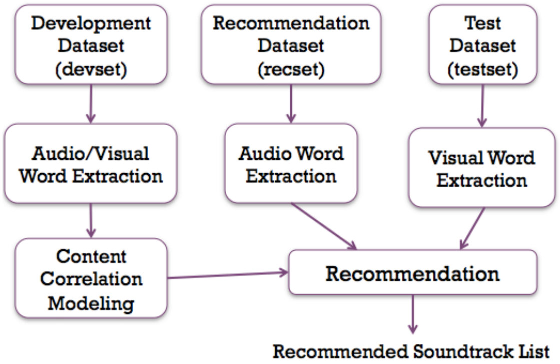
Development Dataset (devset)
Recommendation Dataset (recset)
Test Dataset (testset)
Audio/Visual Word Extraction
Audio Word Extraction
Visual Word Extraction
Content Correlation Modeling
Recommendation
Recommended Soundtrack List

Figure 1: System Architecture of Our Approaches [1].

3. AUDIO WORD EXTRACTION

We use the officially provided audio features including Beat, Key, MFCC, BLF, and PS09 [4] and transform into audio words by discretization or vector quantization (VQ).

For one-dimensional descriptors such as the descriptors of Beat, the equal frequency binning is employed for discretization. The number of bins is set to 19, which is the square root of the number of *devset* [7]. For the multidimensional descriptor, clustering-based vector quantization is performed to group descriptors in the feature space into clusters. For the descriptors of BLF, we use Manhattan distance to measure the distance and utilize the average link and complete link respectively. For the descriptors of PS09 and the FP descriptor of MFCC, we use the Euclidean distance along with the K-means. For each of the three descriptors of MFCC, Gaussian Mixture Model is utilized to model the frame-based representation of an audio. Then K-L divergence along with Earth Mover distance is used to measure the distance, followed by average link and complete link clustering algorithms.

After vector quantization/discretization, each cluster/bin may be regarded as an **audio word** that represents the descriptor belonging to that cluster/bin. An audio descriptor is encoded into an **audio word vector** by the index of the cluster/bin to which it belongs. An audio word vector contains the presence or absence information of each audio word in the soundtrack while the **audio feature vector** for a soundtrack is formed by the concatenation of the audio word vectors respective to all types of descriptors.

4. VISUAL WORD EXTRACTION

The officially provided visual features are based on MPEG-7. In MPEG, the determination of frame types (I, P, B-frames) depends on the compression algorithm of the MPEG encoder. While I-frames may not be key-frames, in our work, the visual features are extracted in the shot-level where the shot boundary detection is based on calculating edge change fraction in temporal domain [8]. Then we extract 13 types of visual descriptors including the color energy, saturation proportion, angular second moment, contrast, correlation, dissimilarity, entropy, homogeneity, GLCM mean, GLCM variance, light median, shadow proportion and visual excitement [1]. Since each of the 13 visual descriptors is scalar, equal frequency binning is performed for generation of **visual words**. **Visual word vectors** and **visual feature vectors** are encoded in the same way as audio word vectors and audio feature vectors.

5. CONTENT CORRELATION MODELING & RECOMMENDATION

We investigate two approaches for learning correlation between audio and visual contents from *devset*.

5.1 CFA (Cross-Modal Factor Analysis)

CFA tries to find the correlation by transforming the audio and visual contents into a common space [2]. Given an **audio feature matrix** X and a **video feature matrix** Y where each row corresponds to the feature vector of a commercial, CFA finds the orthonormal transformation matrices A and B that minimize $||XA\text{-}YB||^2$ where $||M||$ is the Frobenius norm of matrix M. Matrices A and B can be obtained by Singular Value Decomposition (SVD) on X^TY such that $A=U_{xy}$, $B=V_{xy}$, where $X^TY = U_{xy}S_{xy}V_{xy}$. Matrices A and B encode the correlation information. In our work, given a test video f with visual feature vector y_f and a soundtrack m with audio feature vector x_m, the distance $d(m, f)$ between m and f is the Euclidean distance between x_mA and y_fB. The nearest five soundtracks in *recset* are recommended for each test video.

5.2 MLSA (Multi-type Latent Semantic Analysis)

The other approach we adopted is MLSA that exploits pairwise co-occurrence correlations among multiple types of entities (descriptors). MLSA represents the entities and correlations by a unified co-occurrence matrix

$$C = \begin{bmatrix} 0 & M_{12} & \cdots & M_{1N} \\ M_{21} & 0 & \cdots & M_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ M_{N1} & M_{N2} & \cdots & 0 \end{bmatrix}$$

C is composed of $N\times N$ correlation matrices, where N is the total number of descriptor types. M_{ij} is the co-occurrence matrix of descriptor type i and j. C can be decomposed by eigen decomposition. The top k eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ and the corresponding eigenvectors $[e_1, e_2, ..., e_k]$ can span a k-dimensional latent space, which can be represented as an matrix $C_k = [\lambda_1{\cdot}e_1, \lambda_2{\cdot}e_2, \ldots, \lambda_g{\cdot}e_k]$. Given a test video f with feature vector y_f, we first generate the query vector y_q by concatenating y_f with zero audio feature vector. To project onto the latent space, y_q is multiplied by C_k. The likelihood of occurrence $l(a,f)$ between an audio descriptor a and the test video f is the cosine similarity between y_qC_k and the row vector of C_k corresponding to the audio descriptor a. Then the similarity score between a sound track m and the test video f

$$r(m,f) = \textstyle\sum_{\forall\, a \in m} l(a,f).$$

The top five soundtracks in *recset* are recommended for each test video.

6. PERFORMANCE EVALUATION

We take five-fold cross-validation on the *devset* to evaluate the performance of our approach and select the best three models to obtain the ranking result. The original soundtrack of the commercial is regarded as the ground truth and is ranked along with music objects in *recset*. The accuracy in our work is defined as $1\text{-}(rank(g)\text{-}1)/(|C|+1)$ where *rank(g)* is the rank of the ground truth, $|C|$ is the number of music in *recset*. Results with top-2 accuracy for CFA and top-1 accuracy for MLSA are submitted. Table 1 shows the adopted learning algorithms, parameters, accuracy, and the officially rated score of our submitted three results.

Table 1. Performance and Parameters of Submitted Results.

Algorithm	CFA	CFA	MLSA
No. Clusters(GMM, MFCC)	10	10	10
No. Clusters(KL, MFCC)	10	10	10
No. Clusters(FP)	20	20	10
No. Clusters (BLF)	30	10	20
Eigen-number	200	150	400
Accuracy	0.670	0.673	0.547
First rank average	2.292	2.289	2.272
Top-five average	2.264	2.259	2.211

REFERENCES

[1] F. F. Kuo, M. K. Shan, and S. Y. Lee, Background Music Recommendation for Video Based on Multimodal Latent Semantic Analysis, *IEEE Intl. Conf. on Multimedia and Expo,* 2013.

[2] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, Multimedia Content Processing through Cross-Modal Association, *ACM Intl. Conf. on Multimedia*, 2003.

[3] C. C. S. Liem, M. Larson, and A. Hanjalic, When Music Makes a Scene – Characterizing Music in Multimedia Contexts Via User Scene Descriptions, *Intl. Journal of Multimedia Information Retrieval*, Vol. 2, Issue, 1, 2013.

[4] T. Pohle, D. Schmitzer, M. Schedl, P. Knees, and G. Widmer, On Rhythm and General Music Similarity, *Intl. Symp. for Music Information Retrieval*, 2009.

[5] J. Urbano, and M. Schedl, Minimal Test Collections for Low-Cost Evaluation of Audio Music Similarity and Retrieval Systems, *Intl. Journal of Multimedia Information Retrieval*, 2013.

[6] X. Wang, J. T. Sun, Z. Chen, and C. X. Zhai, Latent Semantic Analysis for Multiple-Type Interrelated Data Objects, *ACM Intl. Conf. on Information Retrieval*, 2013.

[7] Y. Yang and G. Webb, Proportional k-Interval Discretization for Naïve-Bayes Classifiers, *European Conf. on Machine Learning*, 2001.

[8] R. Zabih, J. Miller, and K. Mai, A Feature-based Algorithm for Detecting and Classifying Scene Breaks, *ACM Intl. Conf. on Multimedia*,1995.