Fusing Music and Video Modalities Using Multi-timescale Shared Representations

Bing Xu
Department of Information Engineering
The Chinese University of Hong Kong
xb012@ie.cuhk.edu.hk

Xiaogang Wang
Department of Electronic Engineering
The Chinese University of Hong Kong
xgwang@ee.cuhk.edu.hk

Xiaoou Tang
Department of Information Engineering
The Chinese University of Hong Kong
xtang@ie.cuhk.edu.hk

ABSTRACT

We propose a deep learning architecture to solve the problem of multimodal fusion of multi-timescale temporal data, using music and video parts extracted from Music Videos (MVs) in particular. We capture the correlations between music and video at multiple levels by learning shared feature representations with Deep Belief Networks (DBN). The shared representations combine information from multiple modalities for decision making tasks, and are used to evaluate matching degrees between modalities and to retrieve matched modalities using single or multiple modalities as input. Moreover, we propose a novel deep architecture to handle temporal data at multiple timescales. When processing long sequences with varying length, we propose to extract hierarchical shared representations by concatenating deep representations at different levels, and to perform decision fusion with a feed forward neural network, which takes input from predictions of local and global classifiers trained with shared representations at each level. The effectiveness of our method is demonstrated through MV classification and retrieval.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*

Keywords

deep learning; multimodal fusion; music video matching

1. INTRODUCTION

Music and videos, as two popular types of media, cause different human perceptions with music in auditory and video in vision. However, psychology and cognition studies have shown that the information processing procedures of audio and visual signals by human brains are closely related and integrated [2]. In music videos (MVs), music and videos appear in parallel and complement each other. MV expresses the song by using stories and rhythmic dances to complement music. In this work, we try to explore the correlations between musical and visual parts of MVs, and use deep learning to capture such correlations with extracted shared representation across two modalities.

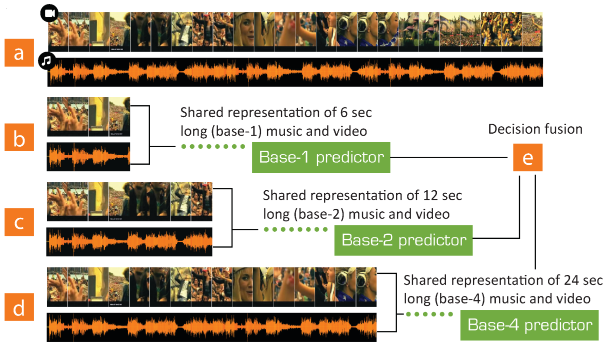
a
b
Shared representation of 6 sec long (base-1) music and video
Base-1 predictor
Decision fusion
e
c
Shared representation of 12 sec long (base-2) music and video
Base-2 predictor
d
Shared representation of 24 sec long (base-4) music and video
Base-4 predictor

Figure 1: Deep learning multi-timescale shared representations. (a) Raw data of a MV segment. (b) 6 sec long base-1 segments. Deep representations for music and video are extracted separately. Shared representations are extracted by concatenating deep representations and input them to another layer. (c) Concatenation of 2 base-1 segments gives 1 base-2 segment. (d) Concatenation of 4 base-1 segments gives 1 base-4 segment. (e) To make a prediction on a 24 sec long time interval, we fuse predictions from 4 adjacent base-1 predictors, 2 adjacent base-2 predictors and 1 base-4 predictor.

One of the commonly recognized matches in MV is music semantics with video content. For example, sad music is paired with sad faces or girls weeping in videos, and exciting party music is paired with dancing scenes and flashing lights. There is existing work on semantically matching music and images with content [10]. The authors built a database with a large number of music-image pairs extracted from MVs, and used Canonical Correlation Analysis (CCA) to capture the relationship between music and image features. However they did not consider any temporal information of music or videos. Since music and images have complex and

MM'14, November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2655069.

distinct structures, they simplified the problem by grouping music-image pairs into five clusters and projected them to the music-attribute space.

In our work, deep learning is used to directly model relationships across modalities instead of simplifying the problem through clustering data pairs as in [10]. And we treat music and videos as temporal data and characterize their dynamic temporal and spacial features. In multimodal data analysis, there have been advances using deep learning approaches [9] [8] [11] [6]. Various kinds of modalities have already been covered by these literatures, including text, audio, music, and image. However, none of the existing work focus on fusing temporal multi-modality data. Temporal data are data that represent states in time, such as music and video. Multimodal problems on temporal data usually concern alignment or synchronization among modalities along time axis, and need to deal with multi-timescale problem, i.e., data sequence may have varying length. Our work provides novel solutions to these issues, demonstrating successful fusion of music and video on multiple level.

Our work makes two key contributions. First, we propose a deep learning framework taking well-crafted inter-frame and intra-frame music and video features as input and fuse them at both semantic and temporal level. This essentially makes sure the model captures semantic matches and takes care of data alignment problem. Second, we propose a novel multi-timescale temporal data fusion approach to deal with variable time length, such that our model can fuse modalities both locally and globally on the time axis. The effectiveness of the proposal multi-modality deep learning fusion approach is demonstrated through two experiments, MV retrieval and classification.

2. DATASET AND FEATURE DESIGN

2.1 Dataset preparation

We obtained 2,000 MVs from YouTube with top rated view counts and acquired 40,000 music-video pairs from them. Most of these MVs were published in 2010 to 2014. The MVs are of diversified styles and were performed by more than 600 singers from different countries.

To speed up feature extraction, all the MVs' video parts are down-sampled to a frame rate of 12fps and resolution is scaled to 320×240 pixels. Audio is downgraded to one channel with a sample rate of 44 kHz.

2.2 Video and music features

We cut music and videos in parallel into small segments of 6 seconds without overlapping or skipping, and extract both local and global features for each segment.

We extract image features within each frame as intra-frame features for videos. They include Histograms of Oriented Gradient (HOG), color histograms, and Local Binary Patterns (LBP). We also design inter-frame features that can capture temporal variation of frames, including histograms of Oriented Optical Flow, distances of HOG, color histograms, and LBP for adjacent frames. We calculate statistics of individual feature sets over all the frames in a segment, namely mean and standard deviation, to capture global characteristics for an entire segment. The final feature dimension for a video segment is 7443.

We use a rich set of mid-level musical features [5], including tempo, attack time, attack slope, spectral characteristics (such as brightness, spread, and roughness), timbre features (such as zero crossing, spectral flux), and tonal features (such as chromagram, key clarity). A pooling step is performed on the raw feature vector with a sliding window along the time axis to average the signal. It computes the mean, median, maximum, and minimum within the window to replace the raw features. Similar as in videos, we calculate statistical values for individual types of musical features over the entire segment, including the mean, standard deviation, slope, period frequency, period amplitude, and period entropy. The final feature dimension for an audio segment is 1052.

3. MULTIMODAL FUSION

3.1 Deep model on single modality

Given a single modality data x with its initial feature representation v_x, we build a multilayer generative deep belief network [4]. A Gaussian-Bernoulli Restricted Boltzmann Machine (RBM) [3] with one layer of latent variables h_x^1 is trained using real-valued v_x as input. Then we treat the activations of h_x^1 as input to train another RBM on top of it, obtaining latent variables h_x^2. More layers of RBM can be further trained in this manner to make the model deep. The whole network can be fine tuned after pre-training steps. Such configuration gives a Deep Belief Network model taking v_x as input and giving deep representation h_x^n as output, where n is the number of hidden layers. Deep representations are less modality-specific than raw data representations, and modeling correlations among deep representations of different modalities is much easier [9] [8]. In our experiments we use two layers of stacked RBMs. And input v_x are normalized to have 0 mean and standard deviation of 1.

3.2 Shared representation for multi-modality

Suppose there are m modalities for data fusion. After obtaining deep representations $h_{x_1}^{n_1}$, $h_{x_2}^{n_2}$, ..., $h_{x_m}^{n_m}$ for each modality x_1, x_2, ..., x_m respectively, where n_i is the depth of DBN for modality x_i, all the deep representations are concatenated into one single vector. Using the merged representation as input, we build another RBM on top of it, which gives hidden layer h as our final shared representation. This is equivalent to learning a joint distribution of all the modalities, and representations of different modalities are fused to one. An graphical illustration of this architecture is shown in Figure 2.

3.2.1 Multi-modality decision making

The process of extracting shared representation for multiple modalities can be treated as preprocessing step. Shared representations capture information from all modalities. Using shared representations in decision making problems such as classification, regression gives better result than directly using single modality data or concatenated multimodality initial feature. Any existing classifiers or regressors can be applied on top of them. We will show the application and experimental result in Section 4.

3.2.2 Multi-modality retrieval

All the layers in our model are undirected. Therefore it is possible to use one or more modalities to reconstruct other modalities. Given only v_p as query, we can search for matched v_q from a database. The hidden representation

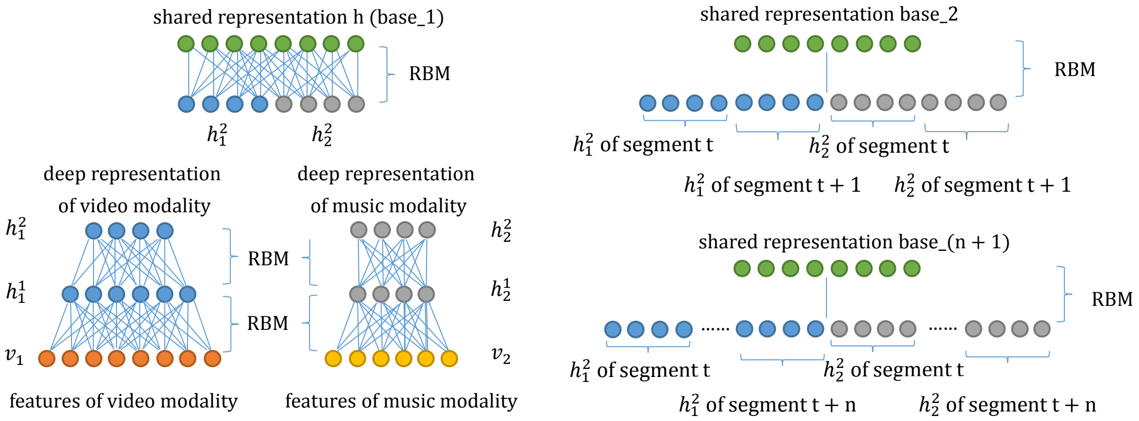

Figure 2: General architecture for extracting multi-timescale shared representation

h_p is first computed from v_p. In the shared RBM, we perform Gibbs sampling by using the conditional distribution $P(h|h_p, h_q)$ to obtain h, and then sample $P(h_q|h)$ to obtain h_q iteratively. Once convergence is reached, h_q can be estimated and easily be backpropagated through its DBN to get $v_q(recon)$. We evaluate the distances between $v_q(recon)$ and all the v_q in the database, and return those with smallest distances as retrieved results.

3.3 Multi-timescale fusion

Decision level fusion is a common strategy in multimodal fusion problems. To deal with long time-scale temporal data, the traditional approach is to fuse modalities on local short segments and average predictions over all short segments in the time interval [1] [7]. However this approach ignores valuable global information that may provide better matching.

Our multi-timescale fusion strategy works in the following way: merging adjacent deep representations within each modality to form hierarchical deep representations (Figure 2), and learning shared representations at different level of hierarchy. It aims at fusing modalities at varying time scales and is much easier to construct than building scale-varying models from bottom up. This approach has several advantages. First, the same set of initial features on the shortest time interval can be reused at different time scales. Therefore, the deep representations for each modality only need to be computed once and the computation is efficient. Second, shared representations at different levels of the hierarchy captures different ranges of cross-modality correlations ranging from local to global, which are complementary to one another. Some long-range characteristics can only be captured by shared representations at higher hierarchy levels.

In a long range matching scenario, one classifier is trained for the shared representation at each hierarchy level . Classifiers from different levels will give decisions on different range (local to global). A final step of decision fusion is performed by using a simple Feed-forward Neural Network (FFNN) to combine the predictions of classifiers at multiple levels.

An exemplar illustration and the architecture of learning multi-timescale shared representations are shown in Figure 1 and 2.

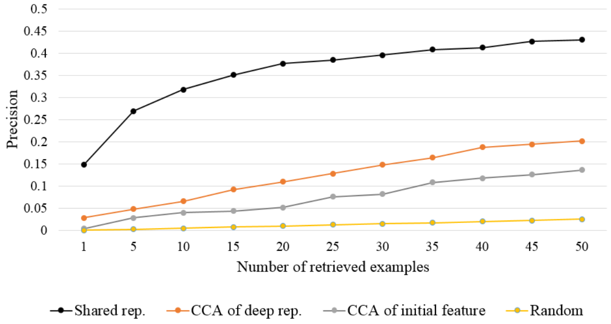

Figure 3: Retrieval precisions using sampling shared rep., CCA with deep rep., CCA with initial features, and random pairing

4. EXPERIMENTS

4.1 Multi-modal retrieval

We construct a video dataset with 2000 video segments, and use 500 audio segments as queries to evaluate the performance of retrieval. Note that the MVs containing these segments are not used in training deep models. We also implement a classic CCA method to evaluate similarity in the same way as [10]. Both CCA and our deep models are trained with 38,000 music-video segment pairs. Figure 3 shows the performance of retrieval using our method and C-CA. We treat the retrieval as successful as long as the top N retrieved examples contains the true match. The precisions of top-N ranks are reported. CCA of deep rep. uses deep representations of videos and audio, and outperforms CCA using initial features as input, which proves deep representations are easier to fuse than initial features. Retrieval with shared representation achieves big improvement over CCA on deep representations or initial features.

4.2 Korean MV classification

A number of YouTube hot clicks are Korean MVs of girl groups and boy bands. Modern Korean MVs are quite distinctive in style. Their music parts have fast tempo and repeated patterns. Their video parts are filled with singers' singing and dancing scenes. In this experiment, we try to learn a classifier to identity Korean MVs. We train an SVM classifier with concatenated initial music and video features

Data input	AP
Video initial feature	0.3123
Video deep rep.	0.5012
Music initial feature	0.2656
Music deep rep.	0.2975
Video and music initial feature (SVM)	0.4542
Shared rep. base-1	0.5585
Shared rep. base-2 (12 sec)	0.5841
Shared rep. base-5 (30 sec)	0.6479
Shared rep. base-10 (60 sec)	0.6025

Table 1: Average precisions of classifiers trained with different data input

as input for comparison. We also demonstrate the effect of multi-timescale fusion of temporal data.

Among 2,000 MVs in our dataset, around 10% are Korean MVs. We cut each MV into segments of 60 seconds, leading to 4,000 segments in total with around 400 Korean MV segments. We randomly obtain 2,000 MV segments containing 200 Korean MV segments as training data, leaving the rest as testing data.

In order to build a multi-timescale fusion model, we cut data into segments of 6 seconds as our base level, denoted by base-1 (40,000 segments in total). Deep representations are extracted for all the base-1 segments. To form a multi-timescale hierarchy, we concatenate the hidden representations of adjacent base-1 segments belonging to the same original 60-second-long segment. We concatenate two adjacent base-1 segments to form 1 base-2 segment (12 seconds long, 20,000 segments in total), concatenate 5 adjacent base-1 segments to form 1 base-5 segment (30 seconds long, 8,000 segments in total), and concatenate 10 adjacent base-1 segments to form 1 base-10 segment (60 seconds long, i.e. the original segments, 4,000 segments in total). One shared RBM is trained on each hierarchy level using base-1, base-2, base-5, and base-10 segments respectively. Then one FFNN classifier is trained with each level's shared RBM individually. We evaluate each classifier's performance using testing data belong to its concatenation level.

With base-1 segments, we also train FFNN classifiers using unimodal input, namely video/music initial features, and video/music deep representations individually. In another comparison, an SVM classifier is trained using concatenated initial music feature and video feature as input. Classifiers' performances are listed in Table 1. Classifiers using shared representations performs much better than those using unimodal representations. Both shared rep. base-1 and SVM use multi-modal input, however, it outperforms SVM by 10%. It shows that the shared representations obtained by deep learning are more effective than the initial features. Noted that AP of shared rep. base-2, base-5 and base-10 are provided for reference only and we do not compare them with others since they use smaller number of data segments (longer length) during training and testing.

In order to deal with long sequences of 60 seconds, we extract predictions from all classifiers and concatenate them into a vector. For each original segment of 60 seconds long, its prediction vector includes 10 local predictions by the base-1 classifier, 5 predictions by the base-2 classifier, 2 predictions by the base-5 classifier, 1 prediction by base-10 classifier. We train a FFNN with prediction vectors and evaluate our multi-timescale fusion model performance. As shown in Table 2, our final classifier (Multi-timescale) gives 0.79 average precision, outperforming the strategy of averaging 10 local predictions using either shared representation or SVM.

Method	AP
Multi-timescale	0.7900
Averaging shared rep. base-1	0.6750
Averaging SVM	0.6130

Table 2: Comparison of multi-timescale representation and averaging on local predictions.

5. CONCLUSIONS

We propose deep models to explore correlations of music and video in MVs. Shared representations for music and videos are learned with DBNs. A retrieval experiment is conducted to demonstrate the power of the shared representations on matched music and videos. We also show that shared representations combines information of music and video and perform better than unimodal representations in the classification task. A novel method is proposed to handle temporal sequences with varying length, by constructing multi-timescale shared representations. It significantly outperforms averaging predictions on local intervals.

6. REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 2010.

[2] R. Cytowic. Synesthesia: A union of the senses. *The MIT Press*, 2002.

[3] G. Hinton. A practical guide to training restricted boltzmann machines. 2010.

[4] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

[5] O. Lartillot and P. Toiviainen. Mir in matlab (ii): A toolbox for musical feature extraction from audio.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. *International Conference on Machine Learning (ICML)*, 2011.

[7] A. Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[8] N. Srivastava and R. Salakhutdinov. Learning representations for multimodal data with deep belief nets. *International Conference on Machine Learning (ICML)*, 2012.

[9] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[10] X. Wu, Y. Qiao, X. Wang, and X. Tang. Bridging music and image: A preliminary study with multiple ranking cca learning. *Proceedings of ACM Multimedia*, 2012.

[11] Z. Yuan, J. Sang, Y. Liu2, and C. Xu. Latent feature learning in social media network. *Proceedings of ACM Multimedia*, 2013.