Noname manuscript No.
(will be inserted by the editor)

Content-oriented Multimedia Document Understanding through Cross-media Correlation

Tong Lu · Yukang Jin · Feng Su · Palaiahnakote Shivakumara · Chew Lim Tan

Received: date / Accepted: date

Abstract This paper presents a novel method for multimedia document content analysis through modeling multimodal data correlations. We hypothesize that the correlation of different modalities from the same data source can help achieve better multimedia content understanding results than one which explores a single modality. We turn this task into two parts: *multimedia data fusion* and *multimodal correlation propagation*. During the first stage, we re-organize the training multimedia data into Modality semAntic Documents (MADs) after extracting quantized multimodal features, and then use multivariate Gaussian distributions to characterize the continuous quantity by latent topic modeling. Model parameters are asymmetrically learned to initialize multimodal correlations in the latent topic space. Accordingly, during the second stage, we construct a Multimodal Correlation Network (MCN) based on the initialized multimodal correlations, and a new mechanism of propagating inter-modality correlations and intra-modality similarities in MCN is further proposed to take the complementary from cross-modalities to facilitate multimedia content analysis. The experimental results of image-audio data retrieval

Tong Lu, Yukang Jin, Feng Su
National Key Laboratory for Novel Software Technology
Dept. of Computer Science and Technology
Nanjing University, Nanjing, China
E-mail: lutong@nju.edu.cn; mg1033012@smail.nju.edu.cn; suf@nju.edu.cn

Palaiahnakote Shivakumara
Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
E-mail: hudempsk@yahoo.com

Chew Lim Tan
School of Computing, National University of Singapore
E-mail: tancl@comp.nus.edu.sg

on a 10-categories dataset and content-oriented web page recommendation on the USTODAY dataset show the effectiveness of our method.

Keywords multimedia documents · multimodal · MAD · MCN · correlation propagation

1 Introduction

Multimedia documents play a wide role in daily life applications by various forms of video streams [1], web pages [2], multimodal corpus [3], multimedia PDF or Microsoft Office documents [4], and even mobile document services like microblogs or social networks composed of photo images, music and textual descriptions [5]. In the past decade, efficient content analysis of multimedia documents (e.g., multimodal data retrieval, multimedia object detection, and content-similar web page recommendation) has drawn much attention from researchers in pattern recognition, multimedia and artificial intelligence. With the production of large data collections favored by cheap digital recording devices and mobile hardware developments, multimedia document analysis has now been a strong commercial demand and is acquiring economic utility in the past years.

Accessing information in multimedia documents is challenging due to the so-called "semantic gap" problem. Correspondingly, the promise of instantaneous semantic access to multimedia document repositories has triggered much attention for methods of document image retrieval [6][7], text detection or recognition [8][9] and video summary [10][11]. Most of these methods follow a similar principle of training a classifier using mono-modal cues such as shape, texture, color, audio, text and motion, as low-level features from the raw multimedia document data. However, multimedia content understanding sometimes requires the usage of several modalities simultaneously rather than depending on a single modality. For example, the presence of a *cat* or a *tiger* would be difficult to directly identify from video streams using visual object recognition techniques since they have similar texture appearances; however, this problem can potentially be easily solved after additionally analyzing their audio characteristics. Thus, combining visual and audio modalities together and even correlating them through a proper mechanism will do more good in understanding multimedia contents. In fact, it has been proved that using mono-modal cues alone in general achieves a fairly low accuracy on unconstrained datasets which consist of multi-modalities [12][13][14]. Therefore, there now exists an urgent need to explore efficient cross-media techniques to facilitate multimedia document content analysis.

Essentially, multimedia data originated from the same source tend to be correlated with each other. It indicates that different modalities in the same multimedia document potentially play a complementary role on solving content analysis tasks. In particular situations, the presence of one modality is even indispensable to understand certain semantics of others. For instance, sound recognition can be a robust compensation of vision-based approaches

in situations from visual obstruction, weak lighting to multi-view observation, especially when the utility of a vision-based application is lost if the visual information is compromised or totally absent. Unfortunately, utilization of multimodal data still faces the following three difficulties. First, low-level features of a single modality, which have a diverse variety of discrete values and a semantic gap towards interpreting high-level contents, are relatively difficult to model multimedia document contents accurately. Second, the traditionally preferred representations of concatenated low-level feature vectors are heterogeneous and cannot be correlated directly. Finally, the high dimensionality of multimodal feature vectors for representing multimedia documents will cause the "curse of dimensionality" problem, making multimedia documents understanding inefficient. As a result, how to correlate heterogeneous low-level features from multi-modalities that have varied dimensions and interpretations is still admittedly a hard problem.

In our previous research, we explored single modality understanding by visual multi-class/multi-view object recognition [15][16], abnormal motion detection [17], and environmental sound event recognition [18] for cross-media information retrieval [19] and movie keyframe extraction [20]. In this paper, we propose a novel multimodal semantic reasoning mechanism for multimedia content analysis with the hypothesis that each modality may compensate for the weakness of the other. We turn the task into two parts: *multimedia data fusion* and *multimodal correlation propagation*. During the first stage, we re-organize the training multimedia documents into Modality semAntic Documents (MADs) after extracting quantized multimodal features. We then use multivariate Gaussian distributions to model the continuous quantity in latent topic modeling. Model parameters are asymmetrically trained to learn multimodal correlations in the latent space accordingly. The number of topics is much smaller than the size of the vocabulary, which leads to efficient computations even for large-scale multimedia datasets that are daily accumulated. During the second stage, a new Multimodal Correlation Network (MCN) is proposed, in which multimodal correlations will be further propagated to take the complementary from cross modalities and accordingly facilitate content analysis of multimedia data.

The main contributions of this paper are as follows:

1. Introduction of a novel framework to capture the inherent multimodal correlations inside multimedia documents. In the framework, MADs will remain the uniformity by mapping multimodal features as occurring frequencies to fuse heterogeneous low-level features. The continuous latent topic modeling avoids the problem of information losing during clustering discrete features and the sensitivity to clustering parameters in latent topic modeling.
2. A scalable MCN representation is proposed to utilize multimodal data relations for interpreting multimedia contents. Moreover, a new mechanism of inter-modality and/or intra-modality propagations on MCN by enhancing multimodal co-occurrences are further presented. Accordingly, the hid-

den relations among multimodal data will be deeply explored and thereby integrated together for content-oriented multimedia understanding in an accurate and robust way.

The experimental results on a 10 categories multimedia dataset and the public USTODAY web page dataset demonstrate the effectiveness of our method in understanding multimedia contents.

The rest of this paper is organized as follows. Section 2 gives a brief overview on multimedia document analysis. Section 3 introduces our framework to explore the inherent multimodal correlations inside multimedia documents. Multimodal correlations are then built and propagated in Section 4. The experimental results and discussions are given in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

The state-of-the-art techniques of multimedia document content understanding can be roughly classified into three categories: multimedia semantic annotation, cross-media correlation modeling, and multimodal data fusion.

The *multimedia semantic annotation* approach bridges the semantic gap between multimedia content and low-level descriptors by providing a semantic annotation framework for describing and representing knowledge both about the content domain and the characteristics of multimedia data. For instance, Jourdan and Bes [22] automatically generate a dynamic multimedia document adapted to the needs of the user from a database of eXtensible Markup Language (XML) fragments for video, images and paragraphs through parameter selecting. Scherp [23] personalizes a coherent multimedia presentation document which reflects the user profile information with semantically-rich multimedia content. [24] and [25] focus on automatically labeling un-annotated multimedia data using textual models. They first represent a visual or sound feature cluster with a dictionary index, and then construct a linked representation to obtain image-text, audio-text and other cross-media translation results. However, despite its success, their methods still suffer from several weaknesses. First, representing each local visual or sound feature by a dictionary index can result in severe loss of information. Second, cross-media index actually focuses on the annotation problem, ignoring semantics reasoning among multimodal data. Sidhom and David [26] further attach annotation defined as textual, graphic or sound to multimedia document source in the context of natural language processing, automatic indexing and knowledge representation. Recently, semantic web techniques have also been adopted for interpreting multimedia contents [27]. For instance, Weiss *et al.* [28] allow for linking low-level MPEG-7 descriptors to conventional Semantic Web ontologies and annotations. Similarly, Mitschick [29] organizes heterogeneous multimedia items and their context through semantic knowledge with the help of semantic web technologies. Saathoff and Scherp [30] propose the Multimedia Metadata Ontology (M3O) for annotating rich multimedia presentations in the web. It

can be easily integrated with multimedia formats such as the W3C standards SMIL and SVG. Unfortunately, the ontology is difficult to define in a uniform way and thereby still remains a long way to go, especially towards automatic multimedia document content understanding.

Generally, the *cross-media correlation modeling* approach [31][32] explores statistic relationship between cross modalities. Yamamoto *et al.* [33] present a picture scheme on utilizing media towards understanding of multimedia documents. After extracting visual and sound features, the known Canonical Correlation (CC) is computed between the feature matrices to learn their correlation [34] and accordingly a hierarchical manifold space can be calculated to make the correlations more accurate [35]. Zhuang *et al.* [13] use transductive learning to mine the semantic correlations among media objects of different modalities so that to achieve cross-media retrieval, by which the query examples and the returned results can be of different modalities, e.g., to query images by an example of audio in multimedia documents. Iria and Magalhaẽs [36] exploit cross-media correlations in the categorization of multimedia web page documents by converting every document into a canonical document-graph representation. Similarly, Wang *et al.* [37] present an iterative similarity propagation approach to explore the inter-relationships between web images and their textual annotations. They first consider web images as one type of objects and their surrounding texts as another, and then construct their links structure via web page analysis to iteratively reinforce the similarities between visual images. However, difficulties still exist due to the heterogeneous feature space and the non-corresponding visual or textual contexts.

The cross-media correlation modeling approach now has many interesting applications. For instance, Wang *et al.* [38] propose an automatic approach for personalized music sports video document generation by using multimodal feature analysis to detect the semantics of events. In [39], Lu *et al.* investigate how to integrate multimodal features for story boundary detection in broadcast news documents. They use a diverse collection of features from text, audio and video modalities and thereby formulate the detection problem as a classification task on the multimodal features. Poignant *et al.* [40] present a video Optical Character Recognition (OCR) system that detects and recognizes overlaid texts in video as well as its application to person identification in video documents. Zhu *et al.* [41] propose a multimodal approach for content based structure analysis of Karaoke music. They find a video text analysis technique to extract the bitmaps of lyrics text from video frames and track the time of its color changes that are synchronized to the singing voice. Theoretically, it can be found that cross-media analysis is indispensable in these multimedia document understanding applications.

Data fusion is another intuitive approach that combines the results of several mono-modality analysis procedures for multimodal analysis and semantic interpretation. Instead of the methods focusing on rule-based combination, Snoek *et al.* [42] identify two general fusion strategies for semantic video analysis, namely, the early fusion and the late fusion, differing in the way they integrate the results from feature extraction on the various modalities. The early

fusion technique combines the extracted features into a single representation as a combination of unimodal features before learning model parameters. Instead, the late fusion technique learns semantic concepts directly from multimodal features and then integrates the learned concept scores to obtain the final semantics. An example of late fusion is in [43], where generative probabilistic models are first learned from visual and textual modalities separatively, and then the learned scores are combined to yield a final detection score for multimedia retrieval. Zhu *et al.* [44] propose a multimodal fusion framework using visual cues and texts for image categorization. Recently, Karaoglu *et al.* [12] evaluate multi-modal object recognition based on visual features fused with text recognition. The detected texts are converted to characters and words, which in turn are used in a text classifier to combine with a bag-of-visual-words image representation. Generally, a disadvantage of data fusion methods is its expensiveness in terms of the learning effort, as every modality requires a separate supervised learning stage. Moreover, the combined representation requires an additional learning stage.

3 The Proposed Framework

Multimedia documents involve multiple information modalities that convey cues related to the nature of the underlying contents, and we concentrate on image and audio here. In this section, we first give the overview of our framework, then represent multimedia documents by the descriptors of the two modalities to provide a uniform representation that is independent of media modalities, and finally introduce our model to learn multimodal correlations.

3.1 Framework Overview

Fig. 1 shows an overview of the proposed framework. Essentially, the framework consists of two parts, namely, *correlation modeling* and *complementarity analysis*, aiming at solving the mentioned difficulties. The purposes of the two stages are respectively as follows:

– In the first stage, the main task is to extract the hidden relations that exist in multimodal data. Directly combining two visual and audio feature vectors together as a new descriptor is unreasonable, which potentially makes multimedia content analysis unreliable. For instance, for a *cat* object, its 128 dimensional visual Scale Invariant Feature Transform (SIFT) feature vector, which is robust to detect in images even under changes of rotation variations, noises and illumination [47], should be in a meaningful way correlated to its 21 dimensional sound Mel Frequency Cepstrum Coefficient (MFCC) feature vector, which is a descriptor of the short-term power spectrum of sound for audio content analysis [48]. For this purpose, we build the relations by using the Probabilistic Latent Semantic Analysis (PLSA)

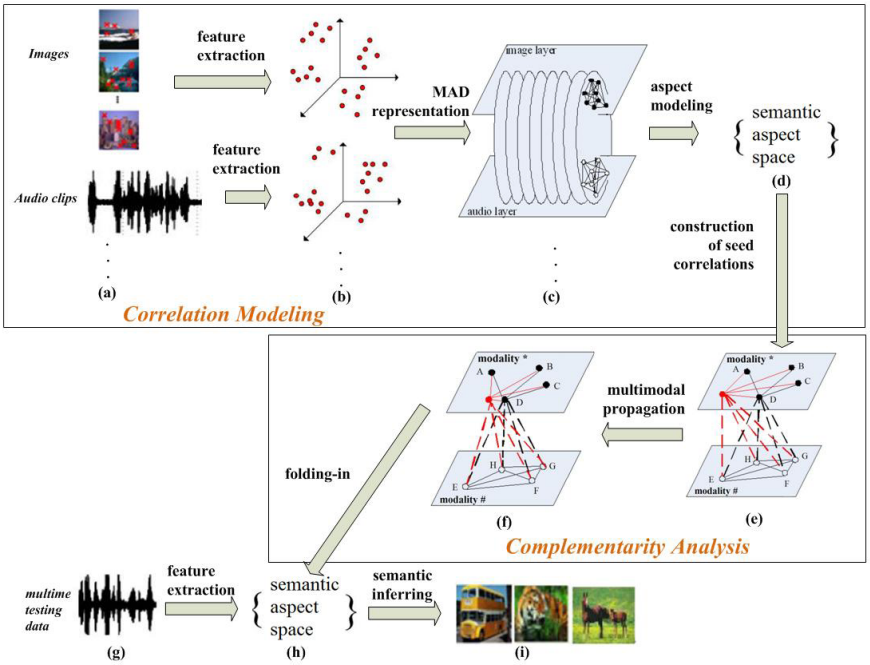


Fig. 1 Overview of the proposed multimodal data content analysis framework: (a) input images and audio clips, (b) extract multimodal features independently, (c) represent multimodal data that have the same semantic annotation into the same MAD in the form of occurrence frequencies of low-level features uniformly, (d) learn the semantic aspect space after topic modeling, (e) initialize MCN, (f) propagate in MCN by contextual co-occurrence modeling, (g) input unknown image or audio clip for interpretation, (h) calculate model parameters for the unknown data based on MCN by the folding-in method, and (i) obtain the final semantic interpretation results.

topic model, which is a statistical technique originally presented for text-based information retrieval [21]. Thereby, the hidden multimodal relations can be initialized by high-level semantics rather than directly by low-level features.

– In the second stage, we integrate all the learned multimodal relations into MCN, in which the dissimilarity from one mono-modality will be further propagated to correspondingly weaken or strengthen another modality. One modality can thereby compensate for the weakness of the other for accurately understanding multimedia contents.

Specifically, during the first stage, we correlate the two different modalities through aspect modeling by 1) inputting annotated image or audio data that are extracted from video streams and collected from websites on Internet (see Fig. 1(a)), 2) independently extracting heterogeneous low-level features (Fig.

1(b)), 3) representing multimodal data that have the same semantic annotation into the same multimedia document in the form of occurrence frequencies of low-level features uniformly (Fig. 1(c)), and 3) learning latent semantic topics by an improved continuous PLSA model (Fig. 1(d)). During the second stage, we initialize the MCN network from all the learned relations that hidden in the training images and audio clips (Fig. 1(e)), then the learned relations are considered as seeds and propagated in the network by contextual co-occurrence modeling (Fig. 1(f)). As a result, multimodal data will be finally correlated with aspect-level interpretations (Fig. 1(h)). When inputting an unknown image or audio clip for testing (Fig. 1(g)), we decide its semantic category by using the learned MCN (Fig. 1(i)). Note that the multimedia data we collected for training and learning can be either structured (e.g., structured sounds of *music* and *speech* due to their formantic or harmonic structure characteristics) or unstructured (e.g., unstructured background sounds of *basketball game* and *soccer* that have a broad noise-like flat spectrum and a diverse variety of signal compositions). Multimedia document content here can thus be interpreted as categorizing the semantics for each unknown image or each audio clip in multimedia documents.

3.2 Representation of Multimedia Documents

We name the multimodal data that have the same annotated semantic category as a *Modality semAntic Document* (MAD). As an example, all the audio tracks of roaring together with all the corresponding visual appearances of *tigers* in the training dataset will be classified into the same MAD. Accordingly, given a collection of MADs which consist of data from visual (V) and audio (A) modalities, we represent them by totally N_{ctg} categories of semantics as

$$MMDocs = \{MAD_1, ..., MAD_c, ..., MAD_{N_{ctg}}\} \tag{1}$$

where an MAD of semantic category c is defined as follows:

$$\begin{aligned} &MAD_c = \\ &\bigcup_{*=V,A} \{ist_i^*(c) | ist_i^* \in \text{category } c, i = 1..IST_c^*\} \end{aligned} \tag{2}$$

$ist_i^*(c)$ denotes a collection of descriptors extracted from the ith multimedia object instance of a specific semantic category c, IST_c^* represents the total number of the instances with the modality $*(* = V or A)$ in category c.

Specifically, for visual modality V, we first extract all the low-level SIFT features from all the IST_c^V visual object instances of category c. Then we use the following vector $Vec(ist_i^V(c))$, which is represented by the occurrence count of every visual feature, to describe the ith visual instance:

$$\begin{aligned} &Vec(ist_i^V(c)) = \\ &\{n(ist_i^V(c), v_1), ..., n(ist_i^V(c), v_p), ..., n(ist_i^V(c), v_{N_c^V})\} \end{aligned} \tag{3}$$

where v_p and N_c^V denote a visual feature and the total feature number extracted from all the visual instances in semantic category c, respectively. $n(ist_i^V(c), v_p)$ is the frequency of v_p that appears in the ith instance.

Similarly, for audio modality A, the jth instance in category c is described by the following vector to avoid the heterogeneous nature brought by modality variations:

$$\begin{aligned} &Vec(ist_j^A(c)) = \\ &\{n(ist_j^A(c), a_1), ..., n(ist_j^A(c), a_q), ..., n(ist_j^A(c), a_{N_c^A})\} \end{aligned} \tag{4}$$

where a_q is an MFCC feature extracted using the same parameters in [18], $n(ist_j^A(c), a_q)$ is its occurring frequency in the jth audio instance, and N_c^A denotes the total sound feature number extracted from all the audio instances in category c.

Followed by the representations, an MAD can be described by the vector $\mathrm{Vec}(MAD_c)$ of dimension $N_c^V + N_c^A$ as follows:

$$\begin{aligned} &\mathrm{Vec}(MAD_c) = \\ &\{n(MAD_c, v_1), ..., n(MAD_c, v_p), ..., n(MAD_c, v_{N_c^V}), \\ &n(MAD_c, a_1), ..., n(MAD_c, a_q), ..., n(MAD_c, a_{N_c^A})\} \end{aligned} \tag{5}$$

where $n(MAD_c, v_p)$ and $n(MAD_c, a_q)$ are respectively defined by

$$n(MAD_c, v_p) = \sum_{ist_i^V(c) \in MAD_c} n(ist_i^V(c), v_p) \tag{6}$$

$$n(MAD_c, a_q) = \sum_{ist_j^A(c) \in MAD_c} n(ist_j^A(c), a_q) \tag{7}$$

As a result, a given training multimedia document dataset is uniformly represented by a set of MADs in the form of an $N_{ctg} \times (N^V + N^A)$ matrix M_{MAD}, where N^V and N^A are the maximums of N_c^V and N_c^A in all the semantic categories, respectively. Since the same feature vector may come from multiple object instances, we additionally store feature vectors and their source object instances for calculating the occurrence counts and the succeeding aspect learning. The proposed MAD representation will remain its uniformity and clarity by mapping multimodal features as occurring frequencies, even when the daily accumulated data in a multimedia document dataset increase sharply. Note that any new modality except the discussed visual and audio modalities can be similarly described in the uniform way. Thereby, it is essentially a general representation to include any number of modalities.

3.3 Construction of the Latent Aspect Space

After showing the representation of multimedia documents, we accordingly turn to modeling multimodal correlations from the training dataset. We use latent aspects that are helpful for content analysis to initialize correlations

among modalities. The latent aspects are modeled based on PLSA, which is a statistical technique for data analysis based on a mixture decomposition derived from a latent class model. The most common application for latent semantic analysis is retrieval from text documents. For a collection of text documents $D = d_1, d_2, ...d_N$ and a vocabulary $W = w_1, w_2, ...w_M$, they can be summarized as an occurrence matrix with terms $c(w_m, d_n)$ representing how many times word w_m appears in document d_n by ignoring the order in which words occur. Specifically, the modeling assumption is that the conditional distributions $P(w|d)$ are approximated by a combination of factors $P(w|z)$ with the mixing weights $P(z|d)$ uniquely defining a point in the latent space. The joint probability model over the documents and words is defined by

$$P(d, w) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \tag{8}$$

It means that for each document d, a latent class is conditionally chosen to the document according to $P(z|d)$, and a word is accordingly generated from that class by $P(w|z)$.

Recently, PLSA has been successfully extended to solve the tasks of unsupervised image annotation [50], the state-of-the-art audio content analysis [51], and even cross-media indexing [52]. Essentially, PLSA models the probability of each occurrence as a mixture of conditionally independent multinomial distributions. The latent variable $z \in Z = z_1, z_2, ...z_K$ in PLSA is also called an *aspect* that represents topics in the text. Accordingly, each document is viewed as a mixture of topics, and each topic is represented by a combination of the words.

However, for multimedia document content analysis, PLSA can not well meet the two mentioned considerations of correlation modeling and complementarity analysis among multi-modalities simultaneously. Specifically, for multimedia content understanding, it still has the following four limitations. First, it is proposed to learn hidden latent aspects from observed variables. We can construct relations among multimodal data through the learned latent aspect distributions indirectly; however, semantic reasoning among different modalities is lacked. Essentially, it is a static probabilistic model, lacking a dynamic correlation analysis mechanism that is required for most multimedia document understanding applications. Second, discrete feature vectors are clustered to construct textual or visual vocabularies for describing documents in PLSA. However, useful information will be inevitable lost during this process. Third, the learning of PLSA parameters are sensitive to the results of feature clustering algorithms. For instance, for the K-mean algorithm, a too large K value will weaken the relations among features, while a too small K value can not well distinguish useful relations instead. It makes PLSA potentially inefficient in computing the optimal clustering parameters for multimedia data in real-life applications. Finally, PLSA is easily disturbed by textual modality annotations. For example, the same image or audio semantic may have several different manually tagged annotations, potentially decreasing the robustness of aspect inferring.

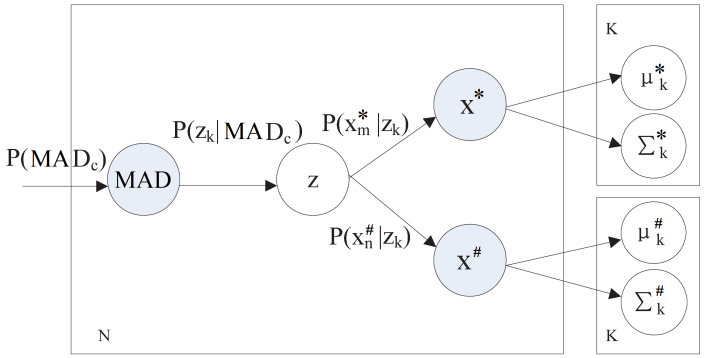


Fig. 2 Multimodal Gaussian PLSA for understanding multimedia documents, where $*$ and # indicate two different modalities.

We propose to use the following stages, namely, 1) modeling continuous PLSA from MADs (Section 3.3), 2) learning model parameters to correlate multimodal data to latent aspects (Section 3.4), 3) building multimodal seed correlations (Section 4.1), and 4) propagating multimodal correlations in multimedia documents (Section 4.2) to interpret multimedia contents. A latent aspect $z_k (k \in 1, .., K)$ is introduced in the generative process to correlate each low-level feature vector $x_i^* (i \in 1, ..., N_c^*)$ and the $MAD_c (c \in 1, ..., N_{ctg})$ it exists in, where $*$ stands for V or A. Since information may be lost between the extracted discrete feature-instances versus real-world matching even if N_c^* is large enough [45], x_i^* is sampled from a multivariate Gaussian distribution rather than a multinomial distribution that is used in the classic PLSA for the unobservable variable z_k. Moreover, since x_i^* is directly sampled, feature descriptors will not be clustered to obtain discrete clusters and thereby the robustness of our model is improved. Fig. 2 shows the graphical model, where the modalities are correlated by sharing the same distribution over the latent aspect $P(z_k|MAD_c)$. Accordingly, due to the fact that the multimodal features in MAD_c are in general independently extracted from multimodal streams in different source multimedia documents, the learned multivariate Gaussian model can well predict the low-level feature distributions of unknown multimedia data.

In our model, the joint probability of an observed pair (MAD_c, x_i^*) is defined by

$$\begin{aligned} p(MAD_c, x_i^*) &= \textstyle\sum_{z_k} p(z_k|MAD_c) P(x_i^*|z_k) \\ &= \textstyle\sum_{z_k} p(z_k) p(MAD_c|z_k) p(x_i^*|z_k) \end{aligned} \tag{9}$$

in which, each feature x_i^* is generated from the K Gaussian distributions, and each Gaussian distribution corresponds to one specific latent aspect z_k. For z_k, its conditional probability distribution of x_i^* is

$$p(x_i^*|z_k) = \frac{1}{(2\pi)^{\frac{Dim}{2}}|\Sigma_k|^{\frac{1}{2}}}\exp^{-\frac{1}{2}(x_i^*-\mu_k^*)^T\Sigma_k^{-1}(x_i^*-\mu_k^*)} \tag{10}$$

where Dim is the dimension of x_i^*, μ_k^* and Σ are the mean vector and the covariance matrix of x_i^* belonging to z_k, respectively.

3.4 Learning Model Parameters

We then estimate the unknowns of the multimodal correlation model. In our MAD representation, we uniformly describe multimedia documents by the occurring frequencies of multimodal words. Essentially, a multimedia document can be described by either a combination of the occurring frequencies of several modalities, or that of any single modality. Thereby, we can asymmetrically fuse multimodal features by first constructing a latent space on one modality and then linking it with another one. Since the heterogeneous features in an MAD share the same latent semantics, the asymmetric learning is feasible and gives a better control of the respective influence of each modality in the latent space. The details of our model learning process are given as follows.

First, for each MAD_c, we choose any modality $*$ ($* = V$ or A) to estimate the parameters of $p(z_k)$, $p(MAD_c|z_k)$, μ_k^* and Σ_k^* through the Expectation Maximization (EM), which is an iterative algorithm for finding the maximum likelihood estimates of parameters with unobserved latent variables [49], on the training multimedia document data set. Note that the knowing of μ_k^* and Σ_k^* is equivalent to know the Gaussian distribution of modality $*$. The EM algorithm is based on the likelihood of the observed data given the parameters of the distributions $p(z_k)$, $p(MAD_c|z_k)$, μ_k^* and Σ_k^*, which iteratively searches for the maximum of this likelihood through the following E and M steps:

$$\begin{aligned} &E(L^c) = \sum_c^{N_{ctg}}\sum_m^{N^*} \\ &n(MAD_c, x_m^*)\log\sum_{k=1}^{K} p(MAD_c|z_k)p(x_m^*|z_k)p(z_k) \end{aligned} \tag{11}$$

E-step: Compute the conditional probability distribution of the latent aspect z_k given the observation pair (MAD_c, x_m^*) from the previous estimation of the model parameters:

$$p(z_k|MAD_c, x_m^*) = \frac{p(MAD_c|z_k)p(x_m^*|z_k)p(z_k)}{\sum_{k=1}^{K} p(MAD_c|z_k)p(x_m^*|z_k)p(z_k)} \tag{12}$$

M-step: Update the parameters of $p(z_k)$, $p(MAD_c|z_k)$, μ_k^* and Σ_k^* with the new $p(z_k|MAD_c, x_m^*)$.

Next, based on the parameters estimated from the other modality #, we adopt the folding-in method [21] to infer $\mu_k^{\#}$ and $\Sigma_k^{\#}$ with the aspect distributions of $p(z_k)$ and $p(MAD_c|z_k)$ kept fixed. $p(x_m^*|z_k)$ and $p(x_n^{\#}|z_k)$ can be respectively inferred according to (12) after knowing the model parameters

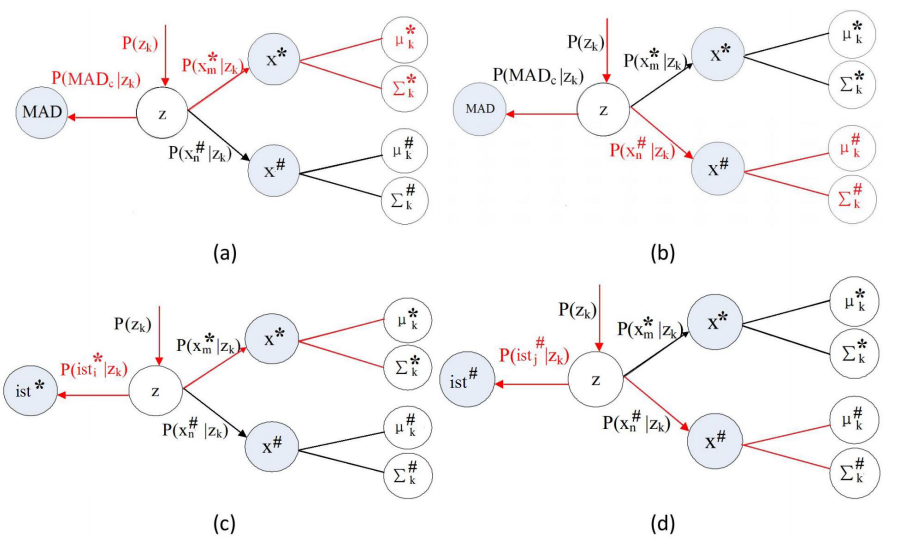


Fig. 3 Asymmetric learning on MADs: (a) learn the distributions of $p(z_k)$, $p(MAD_c|z_k)$, μ_k^* and Σ_k^* from the modality $*$ using the EM algorithm; (b) learn the distributions of $\mu_k^\#$ and $\Sigma_k^\#$ using the folding-in algorithm with the distribution of $p(MAD_c|z_k)$ kept fixed; (c) learn the distribution of $p(ist_i^*|z_k)$ using the folding-in algorithm by fixing the distributions of $p(z_k)$, μ_k^* and Σ_k^*, and (d) similarly learn $p(ist_j^\#|z_k)$ by fixing the distributions of $p(z_k)$, $\mu_k^\#$ and $\Sigma_k^\#$.

of u_k^*, Σ_k^*, $\mu_k^\#$ and $\Sigma_k^\#$. Similarly, the knowing of $\mu_k^\#$ and $\Sigma_k^\#$ is equivalent to know the Gaussian distribution of modality $\#$. Note that the learned parameters remain valid for the multimedia data even out of the training set.

Fig. 3 shows the asymmetric learning process. For every two modalities in MAD_c, the given steps are performed until convergence to learn model parameters.

4 Correlating Multimodal Data in Multimedia Documents

After constructing the aspect model, in this section we first initialize multimodal correlations for constructing MCN and then present our multimodal propagation method.

4.1 Initializing the Multimodal Correlation Network

After learning the feature distributions of the two modalities of $*$ and $\#$, we can in turn estimate the semantic topic distributions of any two multimodal object instances ist_i^* and $ist_j^\#$. By keeping u_k^*, Σ_k^* and the aspect distribution $p(z_k)$ fixed, we once again use the folding-in method to infer $p(ist_i^*|z_k)$, based

on which $p(z_k|ist_i^*)$ is further inferred. $p(ist_j^\#|z_k)$ and $p(z_k|ist_j^\#)$ are learned in the similar way.

Accordingly, the correlations of multimedia object instances will be calculated as follows:

$$\mathrm{Cor}(ist_i^*, ist_j^\#) = \frac{\sum_{z_k} p(z_k|ist_i^*) \cdot p(z_k|ist_j^\#)^{\mathrm{T}}}{\sum_{z_k} |p(z_k|ist_i^*)| \cdot |p(z_k|ist_j^\#)|} \tag{13}$$

where

$$p(z_k|ist_i^*) = \{p(z_{k1}|ist_i^*), ..., p(z_{kp}|ist_i^*), ...\} \tag{14}$$

and

$$p(z_k|ist_j^\#) = \{p(z_{k1}|ist_j^\#), ..., p(z_{kq}|ist_j^\#), ...\} \tag{15}$$

We initialize the correlations among all the multimedia object instances for every pair of modalities in the dataset by asymmetry learning. Suppose any two modalities $*$ and $\#$ are respectively from MAD_{c1} and MAD_{c2}, there are altogether four types of relations among the multimodal object instances, which can be represented by four relation matrixes of $C^{*\#}_{c_1,c_2}$, $C^{\#*}_{c_1,c_2}$, $C^{**}_{c_1,c_2}$ and $C^{\#\#}_{c_1,c_2}$ as follows

$$C^{*\#}_{c_1,c_2} = [\mathrm{Cor}(ist_i^*(c_1), ist_j^\#(c_2))]_{i=1..N^*_{c_1}, j=1..N^\#_{c_2}} \tag{16}$$

$$C^{\#*}_{c_1,c_2} = [\mathrm{Cor}(ist_i^\#(c_1), ist_j^*(c_2))]_{i=1..N^\#_{c_1}, j=1..N^*_{c_2}} \tag{17}$$

$$C^{**}_{c_1,c_2} = [\mathrm{Cor}(ist_i^*(c_1), ist_j^*(c_2)]_{i=1..N^*_{c_1}, j=1..N^*_{c_2}} \tag{18}$$

$$C^{\#\#}_{c_1,c_2} = [\mathrm{Cor}(ist_i^\#(c_1)), ist_j^\#(c_2))]_{i=1..N^\#_{c_1}, j=1..N^\#_{c_2}} \tag{19}$$

The hidden multimodal relations can thereby be initialized accordingly. To illustrate, suppose we need to categorize the new input visual data ist^V_{new} into one of the existing MAD categories that has the most similar semantic. We first infer the aspect distribution of $P(z_k|ist^V_{\mathrm{new}})$ using the folding-in method and learn the model parameters. Then, for each MAD_c, we calculate the relation between ist^V_{new} and the object instances of every modality in MAD_c by (15). As a result, we obtain all the multimodal relations in two forms, namely, *intra similarity* of $\mathrm{Cor}(ist^V_{\mathrm{new}}, ist^V_c)$ and *inter correlation* of $\mathrm{Cor}(ist^V_{\mathrm{new}}, ist^A_c)$. The former indicates the relation of a data pair that belong to the same modality, while the latter reveals the relation from cross modalities. By considering multimedia data as nodes and their relations as edges, multimedia data in the dataset will be represented by the Multimodal Correlation Network (MCN) as shown in Fig. 4, in which each torus implies an MAD, a dotted line connects two nodes of different modalities with the length indicating their *inter correlation* value, while a solid line connects two nodes of the same modality with its length indicating their *intra similarity* value.

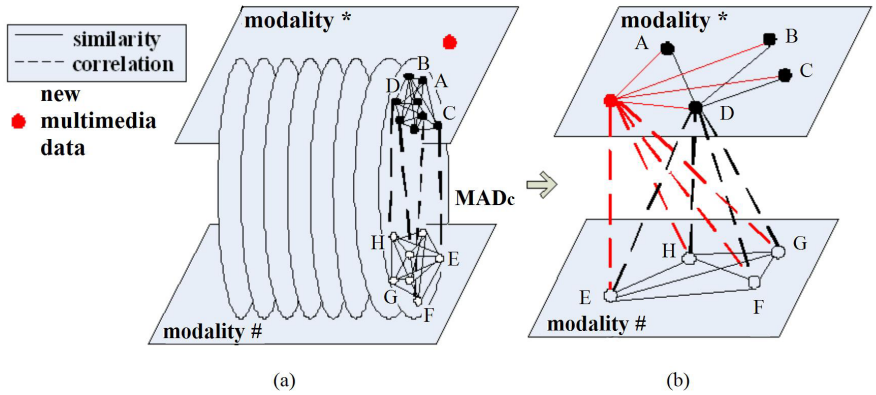


Fig. 4 The Multimodal Correlation Network (MCN) representation for multimedia document content analysis: (a) the MAD representation, (b) the initialization of multimodal correlations among multimedia data. Each torus denotes an MAD, the nodes in which can be from two different modalities (drawn by solid and hollow circles, respectively). Each dotted line connects two nodes which are from different modalities with the length representing an "inter correlation" value, and each solid line connects two nodes which are from the same modality with the length representing an "intra similarity" value.

Thus, for the input example visual data ist^V_{new}, we respectively calculate the intra similarities and inter correlations with the multimedia instances in MAD_c based on MCN as follows:

$$\begin{aligned}&\mathrm{Cor}(ist^V_{\mathrm{new}}, ist^V_c) = \alpha \mathrm{Cor}(ist^V_{\mathrm{new}}, ist^V_c) + \\ &\beta \sum_m \sum_n \\ &(\mathrm{Cor}(ist^V_{\mathrm{new}}, ist^A_m) \cdot C^{AA}_{new,c}(ist^A_m, ist^A_n) \cdot C^{AV}_{new,c}(ist^A_n, ist^V_c))\end{aligned} \tag{20}$$

if

$$\begin{aligned}&\mathrm{Cor}(ist^V_{new}, ist^A_m) > \varepsilon^{VA}, C_{AA}(ist^A_m, ist^A_n) > \varepsilon^{AA}, \\ &C_{AV}(ist^A_n, ist^V_c) > \varepsilon^{AV}\end{aligned} \tag{21}$$

where α and β are coefficients respectively controlling the weight factors, and

$$\begin{aligned}&\mathrm{Cor}(ist^V_{new}, ist^A_c) = \alpha \mathrm{Cor}(ist^V_{new}, ist^A_c) + \\ &\beta \sum_m \sum_n \\ &\mathrm{Cor}(ist^V_{new}, ist^V_m) * C_{VA}(ist^V_m, ist^A_n) * C_{A\bar{V}}(ist^A_n, ist^A_i)\end{aligned} \tag{22}$$

if

$$\begin{aligned}&\mathrm{Cor}(ist^V_{new}, ist^V_m) > \varepsilon^{VV}, C_{V\bar{V}}(ist^V_m, ist^A_n) > \varepsilon^{VA}, \\ &C_{\bar{V}A}(ist^V_m, ist^A_i) > \varepsilon^{AA}\end{aligned} \tag{23}$$

where $\varepsilon^{*\#}_{ij}$ is a multimodal correlation threshold defined as

$$\varepsilon^{*\#}_{ij} = \frac{\sum_{i=1}^{N_{ctg}} \sum_{j \in P} \mathrm{Cor}(ist^*_i, ist^\#_j)}{\sum_i^{N_{ctg}} N_i} = \frac{\sum_{i=1}^{N_{ctg}} \sum_{j \in P} C^{*\#}_{i,j}}{\sum_i^{N_{ctg}} N_i} \tag{24}$$

P is the shortest path connecting with ist_i^* of size N_i.

Accordingly, we construct new correlations between d_{new}^V and the training data set. We finally decide the semantic category of ist_{new}^V as follows:

$$ist_{\text{new}}^* \in \{\text{category}c|\text{Max}(\text{Aver}(MAD_c))\} \tag{25}$$

where

$$\text{Aver}(MAD_c) = \frac{A+B}{\sum_{ist_i^* \in MAD_c}^{I} \sum_j^{\#} 1} \tag{26}$$

$$A = \sum_{ist_i^* \in MAD_c} \text{Cor}(ist_{new}^*, ist_i^*) \tag{27}$$

$$B = \sum_{ist_j^{\#} \in MAD_c} \text{Cor}(ist_{new}^*, ist_j^{\#}) \tag{28}$$

The semantic of any object instance of another modality can be initialized using MCN in a similar way.

4.2 Multimodal Propagation by Co-occurrence Modeling

Essentially, not all the multimodal data in the same dataset can be directly correlated through aspect modeling since there still exist two limitations of accuracy-loss during estimating the distributions and the difficulty in accurately modeling aspect topics. Fortunately, we can further propagate the aspect-level results to obtain a much wider range of multimodal correlations through contextual co-occurrence modeling.

4.2.1 Inter-modality Inferring by Propagating Correlations in MCN

To model co-occurrences, we denote a source multimedia document as $SMD = \{SV, SA\}$, where SV stands for the visual data in it, and SA is its audio data set. Their co-occurrence correlations are characterized by $W = \{W_{ai}\}$, where W_{ai} is a correlation vector between SV and SA. Suppose two document elements of different modalities which are not correlated by aspect modeling, their inter correlation can be constructed and calculated through a propagation way: denoting a correlation between two document elements A and B as $A \leftrightarrow B$, then if $A \leftrightarrow B$ and $B \leftrightarrow C$ are known through aspect modeling, we can further infer a new inter correlation between $A \leftrightarrow C$.

Take Fig. 5 as an example. Suppose an audio element of α_1 and an image element of β_2, which can be from different source multimedia documents, are not correlated by topic-level modeling. Now there are three possible routes for correlating one modality element α_1 to another modality element β_2, which are shown by the dotted arrow lines in Fig. 5. For instance, to propagate along Route 1 of $\alpha_1 \rightarrow \beta_1 \rightarrow \beta_2$, we need to compute the intra-modality similarity between α_1 and β_1 and the inter-modality correlation between β_1 and β_2.

Accordingly, the propagation result of Route 1 is obtained as follows:

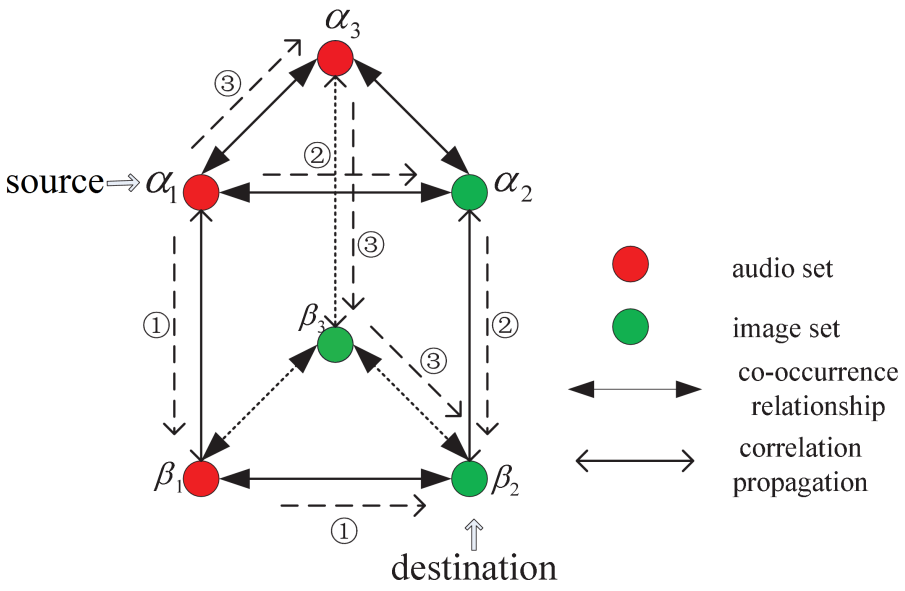


Fig. 5 Propagate inter-modality correlations from α_1 to β_2 by three possible routs of $\alpha_1 \rightarrow \beta_1 \rightarrow \beta_2$, $\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_2$ and $\alpha_1 \rightarrow \alpha_3 \rightarrow \beta_3 \rightarrow \beta_2$.

$$\begin{aligned} &\mathrm{COR}_1^{AV}(\alpha_1, \beta_2) = \\ &\lambda_{11}\mathrm{SIM}^A(\alpha_1, \beta_1) + \lambda_{12}\mathrm{COR}^{AV}(\beta_1, \beta_2) \end{aligned} \tag{29}$$

where $\mathrm{SIM}^*(\alpha_i, \beta_j)$ denotes an intra-modality similarity between the same modality multimedia object instances ($* = V$ or A), while $\mathrm{COR}^{*\#}$ denotes an inter-modality correlation calculated in Section 5.1. λ_{11} and λ_{12} are two normalization coefficients.

Similarly, Route 2 is calculated by $\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_2$ as

$$\begin{aligned} &\mathrm{COR}_2^{AV}(\alpha_1, \beta_2) = \\ &\lambda_{21}\mathrm{COR}^{AV}(\alpha_1, \alpha_2) + \lambda_{22}\mathrm{SIM}^I(\alpha_2, \beta_2) \end{aligned} \tag{30}$$

For Route 3 of $\alpha_1 \rightarrow \alpha_3 \rightarrow \beta_3 \rightarrow \beta_2$, there are two inter correlation paths of $\alpha_1 \rightarrow \alpha_3$ and $\beta_3 \rightarrow \beta_2$ and one intra correlation of $\alpha_3 \rightarrow \beta_3$. Accordingly, the calculation of Route 3 is

$$\begin{aligned} &\mathrm{COR}_3^{AV}(\alpha_1, \beta_2) = \\ &\lambda_{31}\mathrm{SIM}^A(\alpha_1, \alpha_3) + \\ &\qquad \lambda_{32}\mathrm{COR}^{AV}(\alpha_3, \beta_3) + \lambda_{33}\mathrm{SIM}^V(\beta_3, \beta_2) \end{aligned} \tag{31}$$

Finally, the inter-modality correlation between α_1 and β_2 can be calculated by

$$\mathrm{COR}^{AV}(\alpha_1, \beta_2) = max(\mathrm{COR}_1^{AV}, \mathrm{COR}_2^{AV}, \mathrm{COR}_3^{AV}) \tag{32}$$

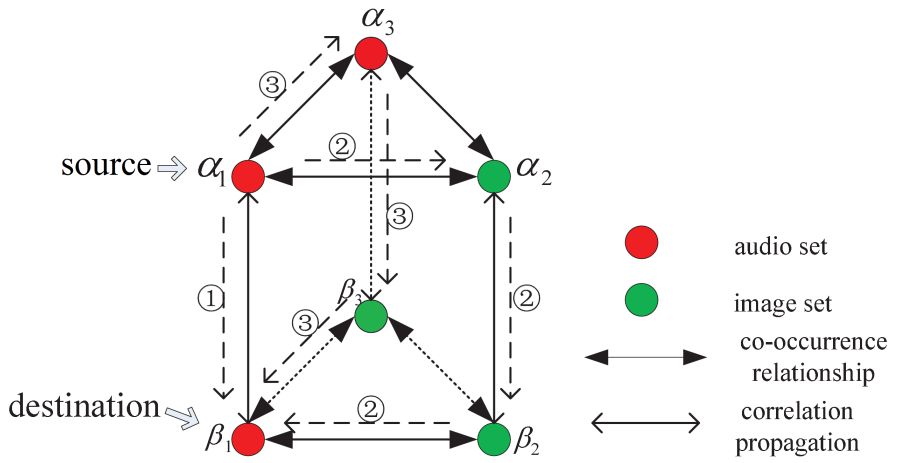


Fig. 6 Enhancement of intra-modality similarities between α_1 and β_1 by three possible multimodal propagation routs of $\alpha_1 \rightarrow \beta_1$, $\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_2 \rightarrow \beta_1$, and $\alpha_1 \rightarrow \alpha_3 \rightarrow \beta_3 \rightarrow \beta_1$.

4.2.2 Re-calculation of Intra-modality Similarities

After propagating all the inter-modality correlations among multimedia documents, we now further enhance intra-modality similarities which are directly calculated by an Euclidean distance measurement. The reason of the enhancement is that an intra-modality similarity directly calculated by discrete feature vector distances is not accurate enough and sometimes even difficult for quantization. For instance in the real-world, sounds are considered in a correlated way to describe a sound scene, e.g., a tight NBA game in general consists of events of whistles, footsteps of people running, broadcasts, and cheers of audience. Unfortunately, whistles and footsteps essentially should not be correlated together directly using distance measures. Similarly, visual features are sometimes difficult to compare especially in situations from visual obstruction, weak lighting to multi-view observation.

We re-calculate intra-modality similarity through contextual co-occurrence modeling as follows. In Fig. 6, suppose we need to re-calculate the intra-modality similarity of document elements α_1 and β_1. Essentially, there exist three different routes to connect α_1 and β_1. The first route is $\alpha_1 \rightarrow \beta_1$, which can be directly expressed by

$$\mathrm{SIM}_1^A(\alpha_1, \beta_1) = \mathrm{SIM}^A(\alpha_1, \beta_1) \tag{33}$$

The second route is $\alpha_1 \rightarrow \alpha_2 \rightarrow \beta_2 \rightarrow \beta_1$. That is, there exist two inter-modality correlations and one intra-modality similarity. Therefore, the calculation of the second route is

$$\begin{aligned} &\mathrm{SIM}_2^A(\alpha_1, \beta_1) = \lambda_{21}\mathrm{COR}^{AV}(\alpha_1, \alpha_2) \\ &+\lambda_{22}\mathrm{SIM}^V(\alpha_2, \beta_2) + \lambda_{23}\mathrm{COR}^{VA}(\beta_2, \beta_1) \end{aligned} \tag{34}$$

Similarly, the third route $\alpha_1 \rightarrow \alpha_3 \rightarrow \beta_3 \rightarrow \beta_1$ can be calculated by

$$\begin{aligned} &\mathrm{SIM}_3^A(\alpha_1, \beta_1) = \lambda_{31}\mathrm{SIM}^A(\alpha_1, \alpha_3) \\ &+\lambda_{32}\mathrm{COR}^{AV}(\alpha_3, \beta_3) + \lambda_{33}\mathrm{COR}^{VA}(\beta_3, \beta_1) \end{aligned} \tag{35}$$

Accordingly, the final intra-modality similarity between α_1 and β_1 can be re-calculated by averaging the different routes as follows

$$\mathrm{SIM}^A(\alpha_1, \beta_1) = ave(\mathrm{SIM}_1^A, \mathrm{SIM}_2^A, \mathrm{SIM}_3^A) \tag{36}$$

Fig. 7 shows an example of the propagated MCN, in which edge length will potentially change after propagations. It is due to the fact that the value of either an inter correlation between two different modalities (see the dotted edges in Fig. 7(a)) or an intra similarity for the same modality (see the solid edges) will be re-calculated. The mentioned example of *cat* and *tiger* that have very similar visual appearances will be well distinguished by additionally propagating the visual similarity into the audio modality. Namely, a long similarity edge that in general indicates a strong intra similarity relation can be weakened and thus the corresponding edge will be shortened as shown in Fig. 7(b). After updating the network for any new input multimedia data, we similarly re-calculate formula (26) to decide the category for it as our semantic interpretation result.

Note that (34) and (38) can be further refined by a set of propagation rules to re-calculate inter-modality correlations or intra-modality similarities (e.g., assigning different values for the coefficients as weights), depending on the characteristics of knowledge-based multimedia contents. It is inspired by the fact that domain knowledge always benefits content analysis in multimedia research, especially for real-life applications. In our previous research, we successfully developed a knowledge-driven system for interpreting real-life drawing documents by turning the complex interpretation process into structural knowledge representation defined as Extended Backus Naur Form (EBNF) and knowledge-driven interpretation on the EBNF-tree [46]. Similarly, by integrating the specific domain knowledge of predefined multimodal relations in the form of rules into the MCN propagation as additional constraints will potentially do help to improve the accuracy of multimodal content interpretations.

Theoretically, the presented method can take any number of modalities, on the condition that the data of a specific modality can be represented by the occurrence frequency of low-level features and be propagated to other modalities by contextual co-occurrence modeling. For example, the motion modality [17] in video can also be integrated by characterizing their spatio-temporal patterns in the form of occurrence frequency (e.g., modeling the

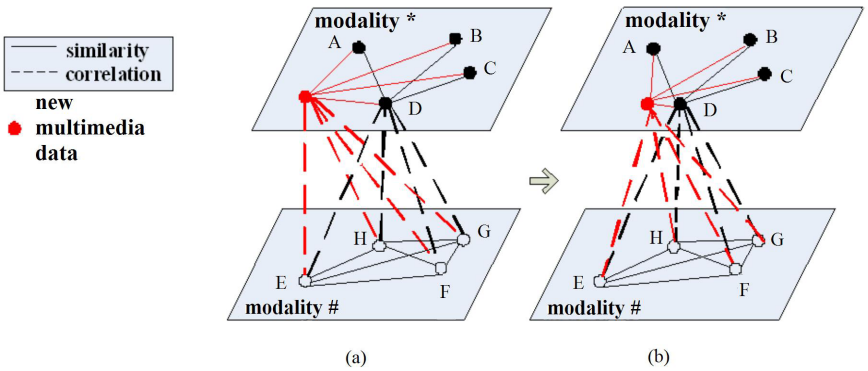


Fig. 7 Multimodal propagation in MCN: (a) Before propagation (see Fig. 4(b)), (b) after propagation. It can be found that the length of either a dotted edge or a solid edge potentially changes during the propagation since the value of an inter correlation or an intra similarity will be re-calculated. Note that the length of a dotted edge essentially corresponds to the value of an inter correlation, while the length of a solid edge similarly corresponds to the value of an intra similarity.

detected anomaly events from video into MADs and correlating them to audio modality events like *scream* for propagations).

5 Experiments and Discussion

To illustrate the effectiveness of our method for content-oriented multimedia document analysis, we conduct our experiments on two applications consisting of *multimedia data retrieval* and *multimedia web page recommendation*.

5.1 Experiment 1: Multimedia Data Retrieval

We collect images and audio clips from the Corel image benchmark dataset [34], the audio dataset [18] and the Internet. Since multimedia content in general covers a wide range, we select 10 categories of Image-Audio data in our first experiment to evaluate the proposed mechanism of multimodal correlation for interpreting multimedia semantics. The dataset consists of *bird*, *car*, *cat*, *elephant*, *explosion*, *horse*, *people*, *tiger*, *thunder* and *water*, each containing 100 sound clips and 150 images. For each category, we select 200 of them as the training data and the rest 50 as the testing data. All the experimental results are with 10-fold cross validation. The selected visual feature is a 128-dimensional SIFT descriptor, while the sound feature is a 21-dimensional MFCC descriptor.

To compare with the existing aspect model in multimedia content analysis, we first perform four groups of experiments of *image-image*, *image-audio*, *audio-audio* and *audio-image* content retrievals by using the classic PLSA.

We randomly select 100 samples from the training set and test set to retrieve their most similar ones, respectively. Since feature clustering is required in PLSA which has the mentioned influences on analyzing the multimodal retrieval performances, we search for the best retrievals by comparing different combinations of K-mean parameters. Table 1 gives the results of the four groups of multimodal content correlating results using the classic PLSA. It can be found that the clustering parameters will bring influences on multimodal content retrieval, and the overall mean Average Precision (mAP) is not so satisfied. mAP is defined as follows:

$$mAP = \frac{\Sigma_{q=1}^{N_q}(AP(q))}{N_q} \tag{37}$$

where

$$AP(q) = \frac{\Sigma_{i \in relevant} precision(i)}{recall(q)} \tag{38}$$

Table 1 Multimodal content retrieval using the classic PLSA. The clustering parameters that are required by PLSA in the experiments are given in the form of K=a,b, where a and b denote the clustering numbers of image and audio data, respectively.

Experiments	K=50,50	K=200,200	K=500,200	K=500,500
image → *image*	0.31	0.53	0.62	0.58
image → *audio*	0.25	0.48	0.55	0.51
audio → *audio*	0.28	0.49	0.65	0.54
audio → *image*	0.23	0.47	0.53	0.50

To illustrate the experiments in Table 1, Fig. 8 shows some multimodal retrieval examples by correlating semantic-related images after inputting audio clips using the classic PLSA. In Experiment 1, we input *cat* and *vehicle* audio clips to retrieve corresponding images. Some of the results are shown in the second row of Fig. 8. It can be found there are wrongly correlated images. For example, the fourth, the sixth and the ninth images actually indicate *tigers* but not *cats*. For comparison, the fourth row of Fig. 8 further shows some image results by respectively inputting *tiger* and *horse* audio clips in Experiment 2. Similarly, it can be found that there exist wrongly correlated images. Note that all the testing audio and image data here have no textual annotations.

We then retrieve multimodal objects using our proposed framework. Fig. 9 shows the mAP results for the 10 categories. It is encouraging that the mAP averagely reaches 65%-78% for multimodal retrieval, showing the effectiveness of the proposed method.

We further compare our model with other multimodal analysis methods of PLSA-WORDS [52] and CCA-based (Canonical Correlation Analysis based) learning [34]. The mAP results are shown in Table 2. PLSA-WORDS is essentially a discrete multimodal model for automatic image annotation, without correlation propagations between multiple modalities. Therefore, it can be seen that PLSA-WORDS does not work well in image-audio retrieval in our experiments. Table 2 also shows that the mean retrieval accuracy of our method

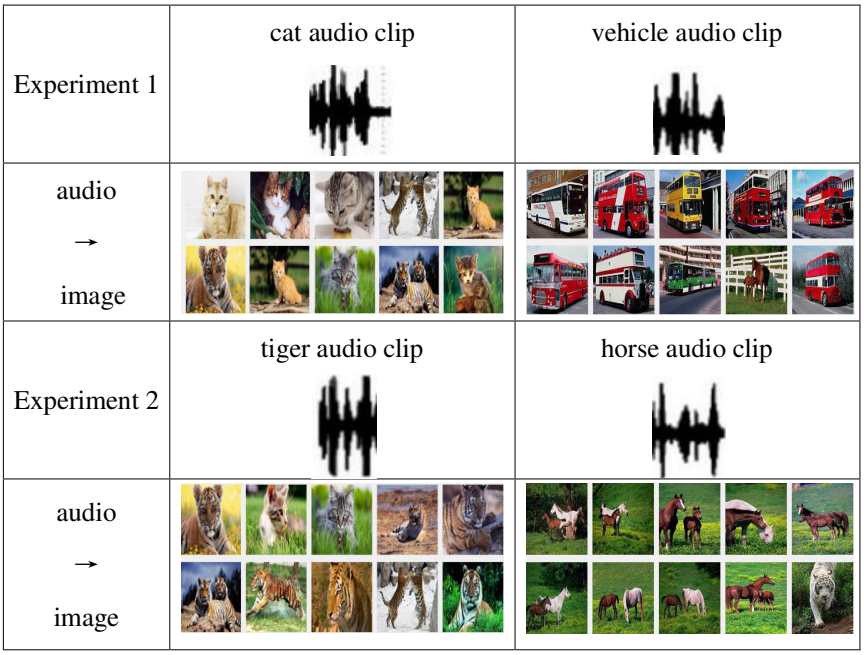


Fig. 8 Multimodal retrieval examples using the classic PLSA, in which the audio clips of *cat*, *vehicle*, *tiger* and *horse* are respectively input in Experiment 1 and Experiment 2 to retrieve the images that have the same semantic interpretation with the input audio data. It can be found that there are wrongly correlated images during both the two experiments using the classic PLSA. Note that all the audio and image data here have no textual annotations.

after multimodal propagation is nearly 17% higher than [34], showing that multimodal data can be better correlated in the aspect space than the low-level feature space. In Table 2, we also note that the average mAP will be effectively improved by propagating the multimodal relations after initializing them using the proposed model (see the mAPs in column "Ours^1" which are before multimodal propagations and the mAPs in column "Ours^2" which are after multimodal propagations for comparisons).

Fig. 10 shows some examples of our multimedia retrieval results on the 10-categories dataset, where the audio-input and image-input retrieval results are respectively shown in Fig. 10(a) and Fig. 10(b). The inputs are approximately 10-sec audio clips of *tiger*, *vehicle* and *horse* in Fig. 10(a), where the most similar 10 results are returned for each semantic category. Comparatively, the multimodal retrieval by using the MCN after multimodal propagations obtains the best results (there are 1 and 3 unexpected multimedia object instances retrieved in the rest two approaches, respectively). In Fig. 10(b), we respectively retrieve three user images of *cat*, *vehicle* and *horse*. The MCN after multimodal propagations again performs the best. It can also be found that in the MCN without multimodal propagations and the CCA-based method,

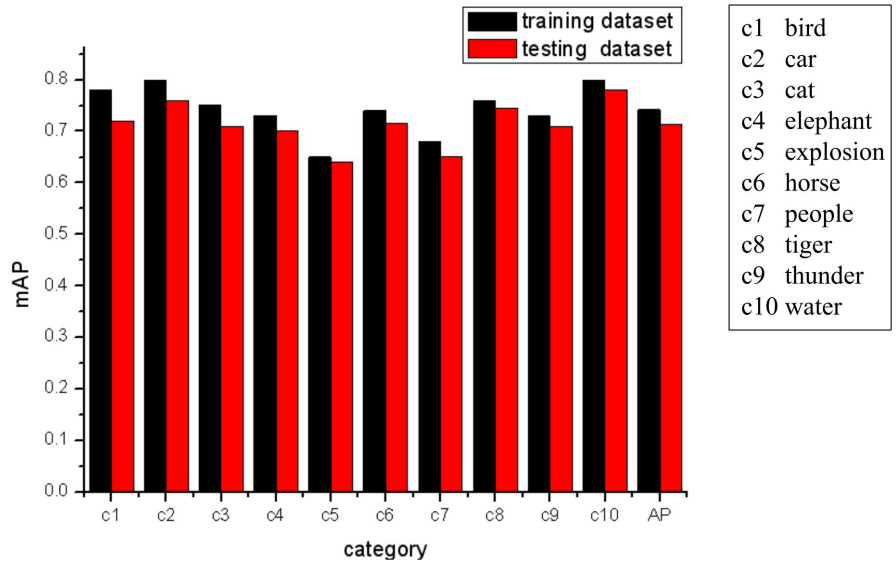


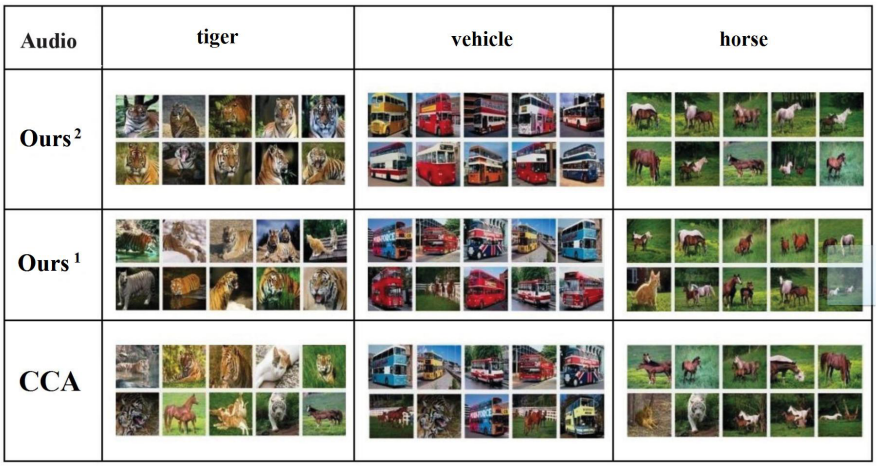
Fig. 9 Experiment 1: mAP of multimedia data retrieval for the 10-categories dataset.

Table 2 Performance evaluation of different models (Ours[1]; before multimodal propagation; Ours[2]: after multimodal propagation). It shows that multimodal data can be better correlated in the aspect space than the low-level feature space. Moreover, the average mAP can be effectively improved by propagating the multimodal relations after initializing them using the proposed model.

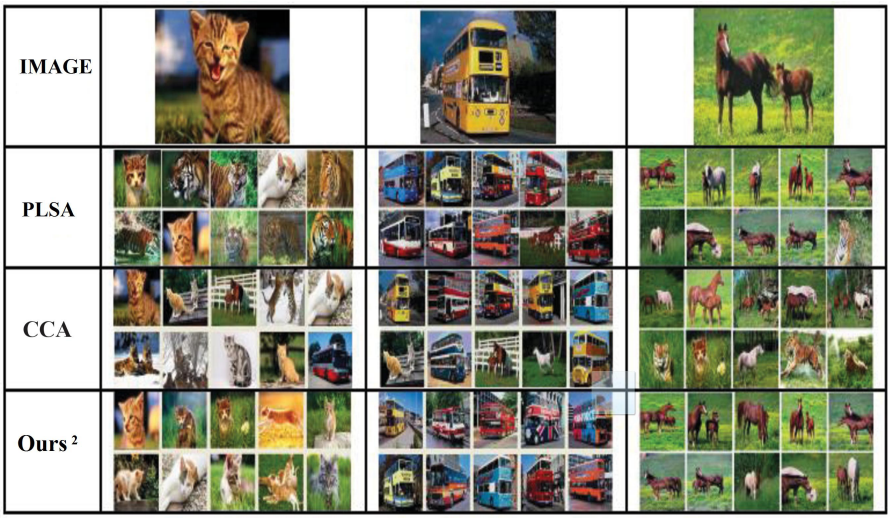
		PLSA	Ours[1]	CCA-based	Ours[2]
$ist_i^I \rightarrow ist_j^I$	path	$ist_i^I \rightarrow ist_j^I$	$ist_i^I \rightarrow MAD_c$	$ist_i^I \rightarrow ..., ist_k^*, ... \rightarrow ist_j^I$	
	mAP	0.62	0.70	0.58	0.75(7.14)
$ist_i^I \rightarrow ist_j^A$	path	$ist_i^I \rightarrow ist_j^A$	$ist_i^I \rightarrow MAD_c$	$ist_i^I \rightarrow ..., ist_k^*, ... \rightarrow ist_j^I$	
	mAP	-	0.67	0.61	0.74
$ist_i^A \rightarrow ist_j^A$	path	$ist_i^A \rightarrow ist_j^A$	$d_i^A \rightarrow MAD_c$	$ist_i^A \rightarrow ..., ist_k^*, ... \rightarrow ist_j^A$	
	mAP	0.65	0.71	0.55	0.76
$ist_i^A \rightarrow ist_j^I$	path	$ist_i^A \rightarrow ist_j^I$	$ist_i^A \rightarrow MAD_c$	$ist_i^A \rightarrow ..., ist_k^*, ... \rightarrow ist_j^I$	
	mAP	-	0.69	0.53	0.72

images of *tiger* and *building* that are visually similar to the input images are returned in the first 10 results. It is obviously hard to distinguish such results without multimodal analysis.

We further evaluate the influence of latent aspect on multimedia retrieval. Fig. 11 shows the mAP values by selecting different numbers of latent aspects from 10 to 80 on different modalities. It can be seen that the multimedia retrieval performance increases sharply before a specific z value and then becomes flat. On our training dataset, approximately $z = 50$ provides a balanced performance between the accuracy and the computational cost in the testing stage. As a comparison, Fig. 12 shows the performance evaluation on latent aspect selection together with clustering parameters of the classic PLSA



(a) Audio retrieval results



(b) Image retrieval results

Fig. 10 Examples of multimedia data retrieval results by inputting sound clips and images, respectively.

method for multimodal analysis. In Fig. 12, we test different combinations of image cluster and audio cluster by (K=50,50), (K=100,100), (K=200,200) and (K=500,200). For each combination, we calculate the mAPs by using 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140 and 150 latent aspects. We can similarly find that there exists such an optimized number of latent aspects for each combination; however, the selection of such optimized aspect number will be easily disturbed by the clustering parameter from any modality. It will

thereby potentially decrease the accuracy of content analysis when analyzing unknown multimedia data.

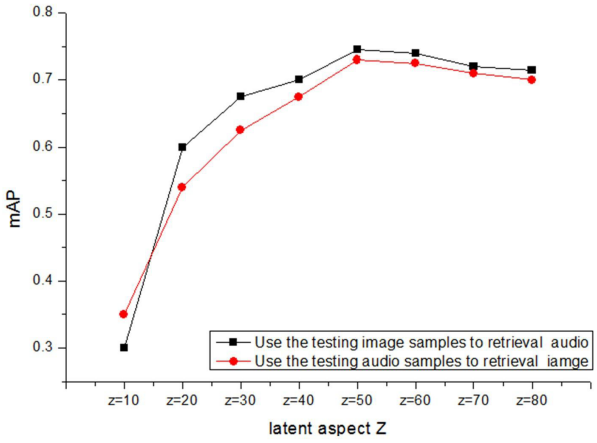


Fig. 11 Influence of the latent aspect z in the proposed model. It can be seen that the multimedia retrieval performance increases sharply before a specific z value and then becomes flat. Note that the proposed model needs not any clustering parameter.

Finally, we illustrate the necessity of asymmetrically fusing multimodal features against directly combining them into a single descriptor for multimedia content analysis using the classic PLSA. In Fig. 13, we compare the mAP with different latent aspect numbers and image/audio clustering parameters of (K=50,50), (K=200,100), (K=200,200), (K=300,200), (K=300,300), and (K=500,200). It can be seen that the asymmetrical fusion of multimodal features averagely achieves a higher mAP against directly combining them together. It is probably because low-level multimodal feature vectors are essentially heterogeneous and should not be combined directly for content analysis.

5.2 Experiment 2: Multimedia content-oriented Web Page Recommendation

Web pages are a popular type of multimedia documents today. Unlike the research in semantic web or social network community by analyzing hyperlink and textual ontology, our multimedia web page recommendation experiment here focuses on exploring content-similar multimedia web page documents by directly comparing the similarities of multimodal data in different web pages. Our dataset is obtained from the USTODAY website via a RSS feed, totally comprising 307 selected web page documents between the 3th of March 2013 and the 12th of May 2013 (altogether nearly 900MB). Each web page document is classified via the USTODAY website by NBA, NHL, TENNIS, NFL, MLB,

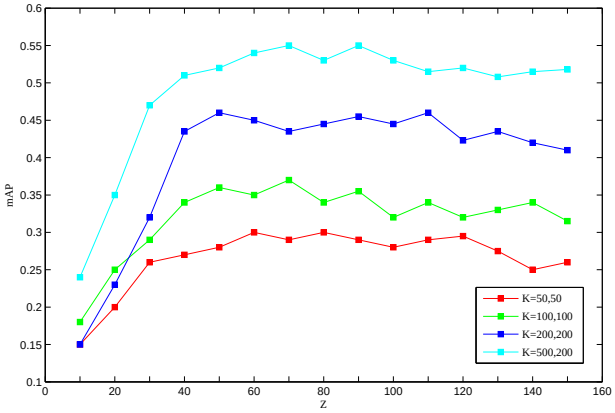


Fig. 12 Influence of the latent aspect z and the clustering parameters in the classic PLSA. We test different combinations of image cluster and audio cluster by (K=50,50), (K=100,100), (K=200,200) and (K=500,200). For each combination, we calculate the mAPs by using 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140 and 150 latent aspects. It can be found that there exists an optimized number of latent aspects for each combination; however, the selection of such optimized aspect number will be easily disturbed by the clustering parameter from any modality.

NCAAF, GOLF, MOVIE, TV, BOOK, GAME, TECHNOLOGY, SOCCER, MUSIC and OTHER via the category assigned by the USTODAY website. For convenience of textual analysis, we manually annotate key words for every web page document, rather than employing a natural language computing algorithm to analyze the textual descriptions in it. Moreover, to simplify the representation of video data in web page documents, we employ our proposed algorithm [20] to extract their visual and sound keyframes, and accordingly divide every video into an image set and a sound set, respectively. Table 3 lists all the categories of the collected multimedia web page document dataset and their corresponding textual annotations.

Suppose P_i is an input web page document, P_i^V stands for its image set and P_i^A is its audio set, we calculate the similarity of P_i and any web page document P_j by summarizing their inter-modality and intra-modality similarities of every multimodal data pair. Then for P_i, its content-similar web pages in the dataset are ordered according to their similarities. Fig. 14 shows the experimental result curve, which is defined by the relationship between the precision and recall of a content-similar searching algorithm, where precision is the percentage of recommended web pages that are relevant, while recall is the percentage of relevant web pages that are recommended. In Fig. 14, we also draw USTODAY's recommendation PR curves, for which we track the "RECOMMENDED FOR YOU" linkage and the "MORE STORIES" linkages on each USTODAY web page. It can be found that the NFL category has the best performance for recommendation; however, the other categories of NBA, NHI and MUSIC are lower than the proposed method. The performances of

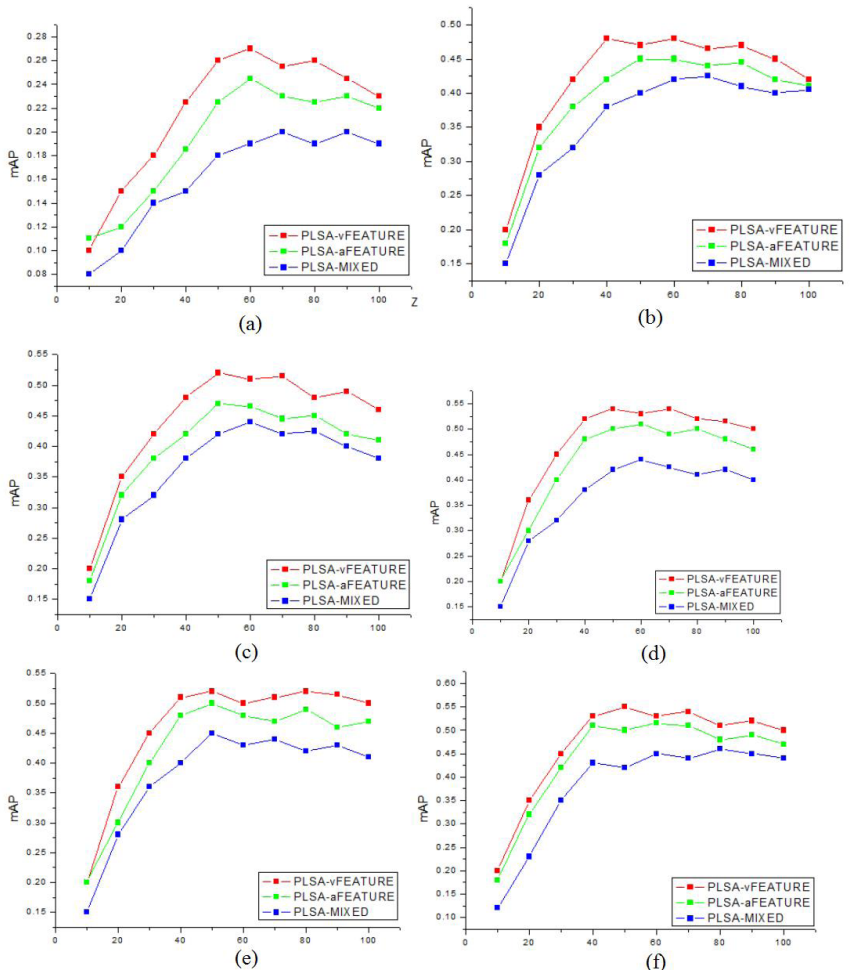


Fig. 13 The comparisons of asymmetrically fusing multimodal features against directly combining the features into a single descriptor for multimedia content analysis using the classic PLSA: (a) K=50,50, (b) K=200, 100, (c) K=200,200, (d) K=300,200, (e)K=300,300, and (f)K=500,200. It can be seen that asymmetrically multimodal feature fusing averagely achieves a higher mAP against directly combining them together.

all the other categories in Table 3 are similar to or even lower than that of NHI, thus their PR curves are not shown in Fig. 14 for simplification. It can be concluded that the "RECOMMENDED FOR YOU" and "MORE STORIES" linkages on USTODAY pages in general focus on particular categories like NFL, rather than automatically recommending other content-similar web pages for most categories.

Fig. 15 shows some examples, where the input web page is shown in the top row of Fig. 15(a), the recommendations of the "MORE STORIES" linkages

Table 3 Multimedia web page document dataset from the USTODAY website.

Category	Pages	Selected textual keywords for annotation
NBA	33	nba, Kobe, James, mvp, play off, finals
NHL	28	nhl, Dan Bylsma, coach, Chicago Blackhawks, Stanley Cup
TENNIS	45	tennis, championships, Djokovic, Davis
NFL	10	nfl, Robert Griffin, Bengals, Pacman Jones
MLB	8	mlb, David Ross, Cory Hahn
NCAAF	13	ncaaf, Paul Myerberg, Kansas, Notre Dame, Alabama, recruit
GOLF	18	golf, Brennan, Tiger-Sergio, Woods, Gavin, Open
MOVIE	13	movie, june, summer, Naomi Watts, trailer,Peter Jackson
TV	17	tv, dance, Daily Show, TV tonight, Major Crimes
BOOK	15	book, Iain Banks, Scottish writer, died, summer, weekend, picks
GAME	22	game, Mario Kart, Nintendo, Destiny, Bungie, E3, guide, news
TECH.	25	technology, Smartphone, cameras, Chuck Leavell, iphone
SOCCER	18	soccer, Barcelona, Neymar, Manchester, Fernandinho, Shakhtar
MUSIC	36	music, Chicago, Dawkins, Koester, West, playlist, rock
OTHER	6	...

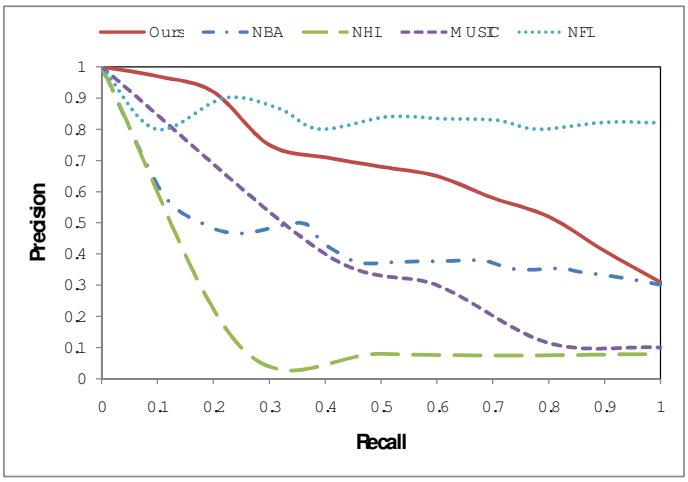


Fig. 14 The PR curve of content-similar web page documents using the proposed method on the USTODAY dataset. USTODAY's PR curves for categories of NBA, NHI, MUSIC and NFL are drawn for comparison. The performances of all the other categories in Table 3 are similar to or even lower than that of NHI, thus their PR curves are not drawn for simplification.

by USTODAY are shown in the middle row of Fig. 15(a), while the bottom row in Fig. 15(a) gives the recommended content-similar web pages using our algorithm. It can be seen that the "MORE STORIES" linkages of USTODAY in general includes a wider range of sport news; however, our recommendation results are mainly in the *NBA* category, which are more content-similar to the input web page document. Note that the layouts of web page documents are simplified for better view. Fig. 15(b) shows more results of web page

(a) Examples by comparing the USTODAY pages and our recommendations.

(b) Content similar web page recommendations by modalities of V and Key-Words, max(V, Key-Words) and the integration of visual and audio modalities.

Fig. 15 Web page content recommendation examples on the USTODAY website.

content recommendation using a single modality or a combination of multiple modalities.

6 Conclusion

This paper proposed a novel multimodal correlation learning and propagation method for multimedia document analysis. We turned the task of correlating multimodal data into two stages consisting of multimedia data fusion

and propagation of multimodal correlation. Multimodal features were extracted from Modality semAntic Documents, allowing us learning and initializing multimodal correlations in the latent topic space by an asymmetrical way. We further propagated the inter-modality correlations and intra-modality similarities using the Multimodal Correlation Network to take the complementary from cross-modalities for facilitating multimedia content analysis. The experimental results showed the effectiveness in retrieving multimodal data. Inspired by our previous work on content interpreting by knowledge representation, our further work includes the improvement of the efficiency for large scale multimedia datasets, and the integration of rule-based propagations by applying domain knowledge in the MCN network for more robust content interpretation.

Acknowledgment

The work described in this paper was supported by the Natural Science Foundation of China under Grant No. 61272218 and No. 61321491, the 973 Program of China under Grant No. 2010CB327903, and the Program for New Century Excellent Talents under NCET-11-0232. The authors thank the anonymous reviewers for their constructive comments, which helped to improve the paper.

References

1. Evangelopoulos G., Zlatintsi A., Skoumas G., Rapantzikos K., Potamianos A., Maragos P., Avrithis Y.: Video event detectio nand summarization using audio, visual and text saliency. IEEE International Conference on ICASSP, pp. 3553-3556 (2009)
2. He J. Y., Weerkamp W., Larson M., Rijke M.: An effective coherence measure to determine topical consistency in user-generated content. International Journal on Document Analysis and Recognition, vol. 12, no. 3, pp. 185-203 (2009)
3. AbdelRaouf A., Higgins C. A., Pridmore T. P., Khalil M. I.: Building a multi-modal Arabic corpus. International Journal on Document Analysis and Recognition, vol. 13, no. 4, pp. 285-302 (2010)
4. Yang Y., Wu F., Xu D., Zhuang Y. T., Chia L. T.: Cross-media retrieval using query dependent search methods. Pattern Recognition, vol. 43, no. 8, pp. 2927-2936 (2010)
5. Erol B., Berker K., Joshi S.: Multimedia clip generation from documents for browsing on mobile devices. IEEE Transactions on Multimedia, vol. 10, no. 5, pp. 711-723 (2008)
6. Lu X. N., Kataria Saurabh, Brouwer W. J., Wang J. Z., Mitra P., Giles C. L.: Automated analysis of images in documents for intelligent document search. International Journal on Document Analysis and Recognition, vol. 12, no. 2, pp. 65-81 (2009)
7. Liang J., DeMenthon D., Doermann D.: Geometric rectification of camera-captured document images. IEEE Trans Pattern Analysis and Machine Intelligence, vol. 30, no. 4, pp. 591-605 (2008)
8. Goto H. Redefining the DCT-based feature for scene text detection. International Journal on Document Analysis and Recognition, vol. 11, pp. 1-8 (2008)
9. Phan T. Q., Shivakumara P., Lu T., Tan C. L.: Recognition of video text through temporal intergration. International Conference on Document Analysis, to appear (2013)
10. Nguyen N. V., Ogier J. M., Charneau F.: PEDIVHANDI: Multimodal indexation and retrieval system for lecture videos. ACCV'12, pp. 382-393 (2012)
11. Peng J., Qin X. L.: Keyframe-based video summary using visual attention clues. IEEE Transactions on MultiMedia, vol. 17, no. 2, pp. 64-73 (2010)
12. Karaoglu S., Gemert J., Gevers T.: Object reading: text recognition for object recognition. ECCV'12, pp. 456-465 (2012)

13. Zhuang Y. T., Yang Y., Wu F.: Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Transactions on Multimedia, vol. 10, pp. 221-229 (2008)
14. Beal M. J., Attias H., Jojic N.: Audio-video sensor fusion with probabilistic graphical models. ECCV, pp. 736-752 (2002)
15. Wang L. M., Wu Y. R., Lu T., Chen K.: Multiclass object detection by combining local appearances and context. ACM Multimedia'11, pp. 1161-1164 (2011)
16. Yin W. C., Lu T., Su F.: A novel multi-view object class detection framework for document image content analysis. International Conference on Document Analysis, Washington, US, pp. 1095-1099 (2013)
17. Ma X. L., Lu T., Xu F. M., Su F.: Anomaly detection with spatic-temporal context using depth images. Internatial Conference on Pattern Recognition, pp. 2590-2593 (2012)
18. Su F., Yang L., Lu T., Wang G.Y.: Environmental sound classification for scene recognition using local discriminant bases and HMM. ACM Multimedia'11, pp. 1389-1392 (2011)
19. Lin W. X., Lu T., Su F.: A novel multi-modal integration and propagation model for cross-Media information retrieval. Multimedia Modeling'12, pp. 740-749 (2012)
20. Jin Y. K., Lu T., Su F.: Movie keyframe retrieval based on cross-media correlation detection and context model. IEA/AIE'12, pp. 816-825 (2012)
21. Hofmann T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, vol. 42, pp. 117-196 (2001)
22. Jourdan M., Bes F.: A new step towards multimedia documents generation. International Conference on Media Futures, pp. 25-28 (2001)
23. Scherp A.: Canonical processes for creating personalized semantically with multimedia presentations. Multimedia Systems, vol. 14, no. 6, pp. 415-425 (2008)
24. Blei D. M., Jordan M. I.: Modeling annotated data. SIGIR, pp. 127-134 (2003)
25. Barnard K., Duygulu P., Forsyth D., Freitas N., Blei D. M., Jordan M. I.: Matching words and pictures. Journal of Machine Learning Research, vol. 3, pp. 1107-1135 (2003)
26. Sidhom S., David A.: Automatic indexing of multimedia documents as a starting point to annotation process. 9th International ISKO Conference on Knowledge Organization for a Global Learning Society (2006)
27. Staab S., Scherp A., Arndt R., Troncy R., Grzegorzek M., Saathoff C., Schenk S., Hardman L.: Semantic Multimedia. In: Reasoning Web. Venis, Italy: Springer (2008)
28. Weiss W., Burger T., Villa Robert, Punitha P., Halb W.: Statement-based semantic annotation of media resources. Proc. of SAMT, pp. 52-64 (2009)
29. Mitschick A.: Ontology-based indexing and contextualization of multimedia documents for personal information management applications. International Journal on Advances in Software, vol. 3, no. 1-2, pp. 31-40 (2010)
30. Saathoff C., Scherp A.: Unlocking the semantics of multimedia presentations in the web with the multimedia metadata ontology. In Proc. of WWW'10, pp. 831-840 (2010)
31. Wang X. J., Ma W. Y., Xue G. R., Li X.: Multi-model similarity propagation and its application for web image retrieval. ACM Multimedia'04, pp. 944-951 (2004)
32. Kyperountas M., Kotropoulos C., Pitas I.: Enhanced Eigen-audioframes for audiovisual scene change detection. IEEE Transactions on Multimedia, vol. 9, no. 4, pp. 785-797 (2007)
33. Yamamoto M., Hikino K., Kijima S., Hirakawa M.: Towards understanding of multimedia documents: a trial of picture book analysis and generation. IEEE International Symposium on Multimedia, pp. 29-36 (2005)
34. Zhang H., Zhuang Y.T., Wu F.: Cross-modal correlation learning for clustering on image-audio dataset. ACM Multimedia'07, pp. 273-276 (2007)
35. Yang Y., Zhuang Y. T., Wu F., Pan Y. H.: Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Transactions on Multimedia, vol. 10, pp. 437-446 (2008)
36. Iria J. and Magalhaẽs J.: Exploiting cross-media correlations in the categorization of multimedia web documents. IJCAI'09 Workshop on Cross-Media Information Access and Mining (2009)
37. Wang J. D., Zeng H. J., Chen Z., Lu H. J., Tao L., Ma W. Y.: ReCoM: reinforcement clustering of multi-type interrelated data objects. SIGIR, pp. 274-281 (2003)
38. Wang J. J., Chng E. S., Xu C. S., Lu H. Q., Tian Q.: Generation of personalized music sports video using multimodal cues. IEEE Transactions on Multimedia, vol. 9, no. 3, pp. 1520-9210 (2007)

39. Lu M. M., Xie L., Fu Z. H., Jiang D. M., Zhang Y. N.: Multimodal feature integration for story boundary detection in broadcast news. ISCSLP, pp. 420-425 (2010)
40. Poignant J., Besacier L., Quenot G., Thollard F.: From text detection in videos to person identification. ICME, pp. 854-859 (2012)
41. Zhu Y., Chen K., Sun Q.: Multimodal content-based structure analysis of Karaoke music. ACM Multimedia'05, pp. 638-647 (2005)
42. Snoek C.G.M., Worring M., Smeulders A.W.M.: Early versus late fuion in semantic video analysis. ACM Multimedia'05, pp. 399-402 (2005)
43. Westerveld T., et al.: A probabilisitc multimedia retrieval model and its evaluation. EURASIP Journal of Application and Signal Process, pp. 186-198 (2003)
44. Zhu Q., Yeh M.C., Cheng K.T.: Multimodal fusion using learned text concepts for image categorization. ACM Multimedia'06, pp. 211-220 (2006)
45. Li. Z. X., Shi Z. P., Liu X., Shi Z. Z.: Automatic image annotation with continous PLSA. ICASSP, pp. 806-809 (2010)
46. Lu T., Tai C.L., Yang H.F., Cai S.J.: A novel knowledge-based system for interpreting complex engineering drawings: theory, representation, and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 8, pp. 1444-1457 (2009)
47. Lowe D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110 (2004)
48. Foote J.: Content-based retrieval of music and audio. Multimedia Storage and Archiving Systems II, Proc. of SPIE, 3229, pp. 138-147 (1997)
49. Carson C., Belongie S., Greenspan H., Malik J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1026-1038 (2002)
50. Monay F., Gatica-Perez D.: PLSA-based image auto-annotation: constrainting the latent space. ACM Multimedia'04, pp. 348-351 (2004)
51. Mesaros A., Heittola T., Klapuri A.P.: Latent semantic analysis in sound event detection. In Proc. of EUSIPCO, pp. 1307-1311 (2011)
52. Monay F., Daniel G. P.: Modeling Semantic Aspects for cross-media image indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pp. 1802-1817 (2007)