FUSING FEATURE AND SIMILARITY FOR MULTIMODAL SEARCH

Guoli Song[1], *Shuhui Wang*[2], *Qi Tian*[3]

[1] University of Chinese Academy of Sciences
[2] Key Lab. of Intell. Info. Process., Inst. of Comput. Tech, Chinese Academy of Sciences
[3] Department of Computer Science, University of Texas at San Antonio

ABSTRACT

It is well known that multiple information fusion can enhance the retrieval performance of multimedia systems. However, what to fuse and how to fuse them are still open issues for multimodal correlation learning. In this paper, we address the problem of combining multiple resources to enhance the multimodal correlation learning ability. We propose two fusion strategies: multi-feature fusion and multi-similarity fusion. For multi-feature fusion, feature concatenation is used to integrate various features. For multi-similarity fusion, three fusion rules are investigated: MIN, MAX, and weighted AVG fusion. The effectiveness of the fusion strategies is evaluated on several state-of-the-art multimodal correlation learning models for cross-modal retrieval tasks. Results suggest that with proper fusion strategy selection, the multimodal retrieval performance can be significantly enhanced.

Index Terms— Multimodal search, data fusion, similarity measure

1. INTRODUCTION

In the real world, data of different modalities, such as text, image and video, often co-exist in a multimedia document to better express the same semantic information. However, the prevailing techniques for searching multimedia information refer to unimodal content modeling. For example, in image retrieval, the similarity between a query image and the corresponding results are retrieved by ranking the similarity between the query and the database images [1]. To facilitate more flexible multimedia retrieval, a retrieving system should be able to carry out a cross-modal retrieval, i.e., the query modality is different from the modality of the retrieved data. To address this problem, researchers devote themselves to developing the multimodal learning methodologies [2, 3, 4, 5]. In most of the existing works, a key issue is how to model correlation between different modalities.

The correlation between multimodal data objects can be described by intra-modality similarity [5] and inter-modality similarity [4, 5]. The inter-modality similarity represents the co-occurrence information or semantic correlation [2] between different modalities. The intra-modality similarity represents the content or semantic similarity within a modality, which plays an important role for multimodal correlation modeling, especially when the inter-modality similarity is noisy. There are various definitions of similarity measures [6], such as Euclidean, Chebyshev, Mahalanobis, Cosine, Hamming distance, etc. An important question is whether these measurements can appropriately model the intra-modality similarity towards multimodal search.

The well-known fusion strategies include early fusion and late fusion [7, 8]. The former fuses multimodal features into a single vector, and the latter fuses multiple modalities in the semantic space. In this paper, we present two fusion strategies: multi-feature fusion and multi-similarity fusion. Similar with the early fusion strategy, multi-feature fusion can be performed by feature concatenation with proper normalization. Multi-similarity fusion combines multiple unimodal similarities derived from different distance measures, which alleviates the curse of dimensionality of feature concatenation. In this paper, we conduct extensive experiments on how better multimodal correlation models can be obtained by adapting the two fusion strategies to several state-of-the-art multimodal correlation learning approaches. The experiment results show that both fusion approaches can enhance the performance of multimodal search system, and better retrieval performance can be achieved by carefully selection of similarity calculation and fusion strategy.

The rest of this paper is organized as follows: Preliminary is introduced in Section 2. The fusion strategy is discussed in Section 3. The experiments are described in Section 4. Conclusion is provided in Section 5.

2. PRELIMINARY

In the first part, we introduce some preliminary knowledge which will be used in this paper, including calculation of sim-

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, National Natural Science Foundation of China: 61025011, 61332016, 61390511 and 61303160 and 863 program of China: 2014AA015202. This work was supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057 and Faculty Research Awards by NEC Laboratories of America. This work was supported in part by National Science Foundation of China (NSFC) 61429201.

978-1-4799-1948-2/15/$31.00 ©2015 IEEE

ilarity and two kinds of data fusion methods.

2.1. Similarity Measures

In information retrieval systems, a common assumption is that similar objects are close to each other in the feature space. The similarity between objects can be measured by computing distances between the feature vectors in the feature space.

Shepard [9] proposed as a universal law that distance and perceived similarity are related via an exponential function:

$$S(x,\ y) = e^{-d(x,\ y)}. \tag{1}$$

Four common distance measures [10], which belong to the Minkowsky family, are presented in Table 1.

Table 1: L_p Minkowski family

Minkowski L_p	$d_p = \sqrt[p]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$
Manhattan L_1	$d_1 = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Euclidean L_2	$d_2 = \sqrt{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^2}$
Chebyshev L_∞	$d_\infty = \max_i \lvert x_i - y_i \rvert$

When we measure similarity between two text documents, cosine similarity is more useful than distance-based measures. Since there are more words that are incommon between two documents, the likelihood that two documents do not share the majority is very high. Word frequencies are represented in a vector. Cosine similarity is then measured as the angle between two vectors, which is represented by [10]:

$$S(x,\ y) = cos(\theta) = \frac{x \cdot y}{\|x\|\ \|y\|}. \tag{2}$$

2.2. Data Fusion Strategies

In this work, the fusion methods are only discussed on visual modality. For multimodal data with multi-lingual text description, a similar discussion can be conducted.

We conduct fusion at feature level and similarity level, which corresponds to two fusion strategies: (1) multi-feature fusion (Fig. 1), (2) multi-similarity fusion (Fig. 2). First, various features are extracted from input image. Then the similarity score can be computed according to the measures mentioned in Section 2.1. Before fusing data, a 2-norm normalization procedure is necessary for both strategies, because different ranges of different feature or similarity values will lead to inappropriate dominance of high dimensional features.

At the feature level, the fusion can be conducted in three ways: concatenation, selection and extraction. The concatenation is simply combining the component feature vectors to a super feature vector. The selection means selecting a subset from all the available features. The extraction refers to project the original features into a subspace so that the new features are a linear or nonlinear combination of the original features. As we can see, only concatenation retains all the original feature vectors, and it is the simplest and easiest to implement. Therefore, we study the vector concatenation for multi-feature fusion in this paper.

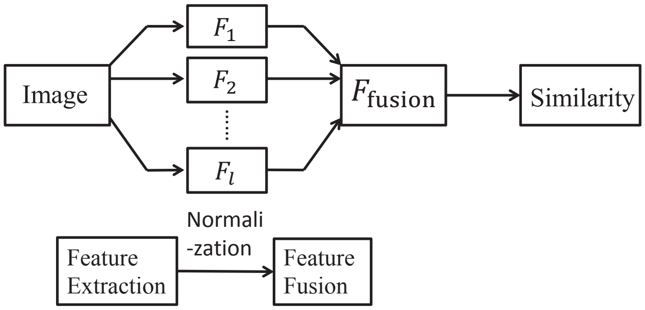

Fig. 1: Multi-feature fusion

Compared with the multi-feature fusion, the similarity level fusion can avoid handling high dimensional feature vectors. In multi-similarity fusion, the fused similarity value can be computed using the following fusion rule:

$$S_{fusion}(x,\ y) = F_{c=1}^{k}\left(S_c\left(x,\ y\right)\right), \tag{3}$$

where $S_c,\ c = 1, \ldots, k$, are different similarity measures between feature vectors x and y, and F is the fusion rule, including MIN, MAX, and weighted AVG fusion. MIN and MAX rules represent the assignment of extreme values of similarity and it is thus hardly surprising that both can be highly sensitive to the presence of “noise”. The weighted AVG rule is expected to be more stable against the noisy points. However, MAX or MIN rule may be more discriminative than weighted AVG rule.

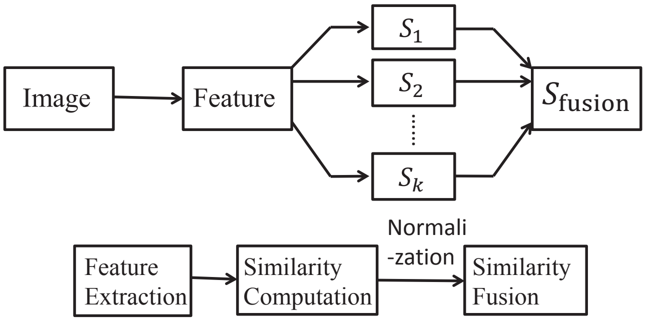

Fig. 2: Multi-similarity fusion

3. IMPLEMENTATION

First, images and texts are represented as vectors in feature spaces $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively, where n_c $(c = 1,\ 2)$, denote the dimensionality of the corresponding modality feature. For convenience, index notation c is introduced to indicate image ($c = 1$) or text ($c = 2$) modality.

3.1. The Problem

Given a text query $T_q \in \mathbb{R}^{n_2}$, the goal of multimodal correlation search is to return the closest match $I_q \in \mathbb{R}^{n_1}$ in the image space, and vice versa.

Let $\mathcal{I} \in \mathbb{R}^{n_1 \times N_1}$ and $\mathcal{T} \in \mathbb{R}^{n_2 \times N_2}$ represent an image set and an text set, respectively, where N_c, ($c = 1,\ 2$), denotes the number of data in the corresponding modality. $S^1 \in \mathbb{R}^{N_1 \times N_1}$ and $S^2 \in \mathbb{R}^{N_2 \times N_2}$ are intra-modality similarity matrices. S^c_{ij} in S^c, ($c = 1,\ 2$), the similarity between the *i*th and *j*th feature vectors, can be derived according to measures mentioned in Section 2.1.

In this work, each text on $\mathbb{R}^{n_2}$ is represented with latent Dirichlet allocation (LDA) model [11]. Cosine similarity is used to measure the relationship between text documents. The construction of image similarity matrices is presented in the following section.

3.2. Multi-feature Fusion

The feature level fusion integrates various features (Figure 1). For each image, we employ three popular visual descriptors: a 128-dimensional SIFT feature vector, a 128-dimensional HOG feature vector, and a 512-dimensional GIST feature vector. SIFT and HOG representation are based on bag-of-words (BOW) model. More precisely, a bag of descriptors is first extracted from each training image. Then a 128-dimensional visual word codebook is learned with K-means clustering algorithm. Descriptors extracted from each image are finally vector quantized with this codebook to produce a vector of visual word counts. These features characterize different aspects of images. We concatenate two or all of them to obtain a fusion representation. Then the image similarity matrices are derived according to equation (4), using the most widely used Euclidean distance.

3.3. Multi-similarity Fusion

In the case of multi-similarity fusion (Figure 2), we compute multiple distance measures between SIFT feature vectors. We employ three common distance measures: Manhattan distance (d_1), Euclidean distance (d_2), and Chebyshev distance (d_∞). The similarity measures are denoted by S^1_1, S^1_2, and S^1_∞, respectively, which are derived by

$$S^1_p = \mathrm{e}^{-d_p^2/2},\ p = 1,\ 2,\ \infty. \tag{4}$$

And three fusion rules are considered: MIN, MAX, and weighted AVG fusion.

Denote the fused similarity matrix as S^1_f. By the weighted AVG rule, S^1_f is the weighted average of some of the above similarity measures:

$$\begin{aligned} &S^1_f = \sum_p w_p S^1_p\,, \\ &\text{s.t.} \left\{ \begin{array}{l} 0 \leqslant w_p \leqslant 1 \\ \sum\limits_p w_p = 1 \end{array} \right. ,\ p \in \Lambda \subseteq \{1, 2, \infty\}\,. \end{aligned} \tag{5}$$

Under the MAX rule, the similarities in S^1_f are computed in the following way:

$$S^1_f\left(i,j\right) = \max_p \mathscr{S}^1_p\left(i,j\right), \tag{6}$$

where

$$\begin{aligned} &\mathscr{S}^1_p\left(i,j\right) = \left\{ \begin{array}{ll} S^1_p\left(i,j\right), & \text{if } S^1_p\left(i,j\right) \geqslant \theta_p, \\ 0 & , \text{ otherwise}, \end{array} \right. \\ &\forall i,j \in \left\{1,\ldots,N_1\right\},\ p \in \Lambda \subseteq \{1,2,\infty\}\,. \end{aligned}$$

We introduce several threshold parameters θ_p to modify the final similarity values, considering the fact that for any given vector x, $\|x\|_\infty \leqslant \|x\|_2 \leqslant \|x\|_1$. Besides, it is helpful to reduce the effect of noise by introducing thresholds. The fused similarity matrices are very sensitive to the value of the thresholds, where a careful parameter tuning is in need. The MIN rule is similar to the MAX, so we omit the discussion due to the limited space.

4. EXPERIMENTS

4.1. Experimental Setup

The well-known Wiki data set [2] is used for evaluation, which contains a total of 2866 documents. These documents are image-text pairs, and annotated with a label from 10 semantic categories. All experiments are implemented for two common cross-modal retrieval tasks: (1) image query vs. text database, (2) text query vs. image database. A random 80/20 split of the data set is used to produce a training set and a testing set.

In our experiments, we take as baseline a multimodal hash function learning method: MLBE [5], which uses 128-dimensional SIFT feature to represent image modality.

Mean Average Precision (MAP) is used to evaluate the retrieval performance. MAP measures whether the retrieved data belong to the same class as the query. The larger the MAP, the better the retrieval performance. Similar as MLBE, the number of retrieved documents is set to 50.

4.2. Results on Multi-feature Fusion

In the first experiment on multi-feature fusion, the resulting MAP measures are provided in Table 2. We can see that the best result is from the concatenation of SIFT and HOG

Table 2: MAP Scores of Multi-feature Fusion

Rule	MLBE			CVH		
	img-query	txt-query	avg	img-query	txt-query	avg
SIFT	0.3553	0.4718	0.4136	**0.2722**	0.3437	0.3080
SIFT+HOG	**0.5467**	**0.5036**	**0.5252**	0.2715	0.3793	0.3254
SIFT+GIST	0.3640	0.4756	0.4198	0.2306	0.4590	0.3448
SIFT+HOG+GIST	0.4922	0.5018	0.4970	0.2351	**0.4972**	**0.3662**

Table 3: MAP Scores of Multi-similarity Fusion

Rule	weighted AVG			MAX			MIN		
	img-query	txt-query	avg	img-query	txt-query	avg	img-query	txt-query	avg
S_2^1	0.3553	0.4718	0.4136	0.3553	0.4718	0.4136	**0.3553**	**0.4718**	**0.4136**
S_1^1-S_2^1	**0.4521**	0.4494	0.4508	0.4456	0.5220	0.4838	0.2611	0.3135	0.2873
S_2^1-S_∞^1	0.3956	0.4966	0.4461	**0.4617**	0.5230	**0.4924**	0.2591	0.3642	0.3322
S_1^1-S_2^1-S_∞^1	0.4066	**0.5486**	**0.4776**	0.4416	**0.5305**	0.4861	0.2778	0.4070	0.3424

features. It shows that combining multiple features can indeed achieve higher accuracy. However, other combinations of features enhance the performance little and there are an unexpected result. Notice that the concatenation of three features doesn't outperform the concatenation of two features. Therefore, selection of feature representation is important for multi-feature fusion, and combination of more features can produce better results, though not always.

To further show that combining multiple features can improve accuracy, we conduct another set of experiments based on CVH [4]. The results are provided in Table 2. Given space limitations, details are omitted here.

4.3. Results on Multi-similarity Fusion

We perform three sets of experiments on multi-similarity fusion. SIFT feature vectors are employed to compute similarity matrices. The first set is conducted using weighted AVG fusion rule. We assign different weighting factors (w_1,w_2,w_∞) to the similarity matrices. The results are shown in Table 3. Take S_1^1-S_2^1 as an illustration, the corresponding MAP scores are derived from the weighted average of similarities computed by Manhattan distance and Euclidean distance. We can see the setting of similarity weights has great impact on the performance. The results show that weighted average of multiple similarity measures has superiority over a single similarity in some contexts. Besides, we can see that retrieval performance not only depends on good feature but also on good similarity, and Chebyshev distance can be taken as an alternative for Euclidean distance to perform cross-modal retrieval.

The second set of experiments is performed on MAX rule, and the third set is on MIN rule. Under these two rules, we need choose appropriate thresholds (θ_1, θ_2, θ_∞), not too big nor too small. In this paper, we use given priori values for these thresholds. The resulting MAP scores are shown in Table 3. On the whole, results for the MAX rule are better than weighted AVG rule. However, results for the MIN rule are very poor, and MIN fusion even has negative impact on the performance.

5. CONCLUSIONS

In this paper, we address the problem of combining multiple resources to enhance the multimodal correlation learning ability. We propose two fusion strategies: multi-feature fusion and multi-similarity fusion. Experimental results show that it is indeed possible to achieve retrieval enhancements with data fusion. The fusion rules have great impact on the results. For both strategies, using all the available information to perform fusion may not be optimal. The performance of MAX rule is constantly better than other rules. In future study, we will conduct theoretical analysis of data fusion, and apply it to the design of fusion methods for multi-modal retrieval.

6. REFERENCES

[1] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, 2006.

[2] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy,

and Nuno Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in *ACM Multimedia*, 2010.

[3] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell, “Learning cross-modality similarity for multinomial data,” in *ICCV*, 2011, pp. 2407–2414.

[4] Shaishav Kumar and Raghavendra Udupa, “Learning hash functions for cross-view similarity search,” in *IJCAI*, 2011.

[5] Yi Zhen and Dit-Yan Yeung, “A probabilistic model for multimodal hash function learning,” in *KDD*, 2012.

[6] Manesh Kokare, BN Chatterji, and PK Biswas, “Comparison of similarity metrics for texture image retrieval,” in *TENCON 2003*. IEEE, 2003, vol. 2, pp. 571–575.

[7] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[8] Cees Snoek, Marcel Worring, and Arnold W. M. Smeulders, “Early versus late fusion in semantic video analysis,” in *ACM Multimedia*, 2005, pp. 399–402.

[9] Roger N Shepard, “Toward a universal law of generalization for psychological science,” *Science*, vol. 237, no. 4820, pp. 1317–1323, 1987.

[10] Sung-Hyuk Cha, “Comprehensive survey on distance/similarity measures between probability density functions,” *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.