

Received September 21, 2018, accepted October 26, 2018, date of publication October 31, 2018, date of current version December 3, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2878868

# A Benchmark Dataset and Learning High-Level Semantic Embeddings of Multimedia for Cross-Media Retrieval

SADAQAT UR REHMAN<sup>1</sup>, (Student Member, IEEE), SHANSHAN TU<sup>2</sup>, (Member, IEEE),  
YONGFENG HUANG<sup>1</sup>, (Senior Member, IEEE), AND OBAID UR REHMAN<sup>3</sup>

<sup>1</sup>Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China

<sup>2</sup>Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

<sup>3</sup>Sarhad University of Science and Information Technology, Peshawar 25000, Pakistan

Corresponding author: Shanshan Tu (sstu@bjut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U1705261, Grant U1405254, Grant U1536115, and Grant U1536207, and in part by the National Key Research and Development Program of China under Grant 2016YFB0801301.

**ABSTRACT** The selection of semantic concepts for modal construction and data collection remains an open research issue. It is highly demanding to choose good multimedia concepts with small semantic gaps to facilitate the work of cross-media system developers. However, very little work has been done in this area. This paper contributes a new, real-world web image dataset for cross-media retrieval called FB5K. The proposed FB5K dataset contains the following attributes: 1) 5130 images crawled from Facebook; 2) images that are categorized according to users' feelings; 3) images independent of text and language rather than using feelings for search. Furthermore, we propose a novel approach through the use of Optical Character Recognition and explicit incorporation of high-level semantic information. We comprehensively compute the performance of four different subspace-learning methods and three modified versions of the Correspondence Auto Encoder, alongside numerous text features and similarity measurements comparing Wikipedia, Flickr30k, and FB5K. To check the characteristics of FB5K, we propose a semantic-based cross-media retrieval method. To accomplish cross-media retrieval, we introduced a new similarity measurement in the embedded space, which significantly improved system performance compared with the conventional Euclidean distance. Our experimental results demonstrated the efficiency of the proposed retrieval method on three different datasets to simplify and improve general image retrieval.

**INDEX TERMS** Cross-media retrieval, FB5K dataset, high-level semantic embeddings.

## I. INTRODUCTION

The current era has seen rapid growth in Multimedia Information Retrieval (MIR). Despite constant hard work in the development and construction of new MIR techniques and datasets, the semantic gap between images and high-level concepts remains high. We need a promising model to focus on modeling high-level semantic concepts, either by image annotation or by object recognition to diminish this semantic gap. Numerous real-world methods [1]–[7] have been introduced for this kind of concept-based multimedia search system. Among several of these methodologies, the first step is dataset selection for high-level concepts and small semantic gaps, which are relatively easy for machine understanding and training

Work on cross-media retrieval can be categorized in to two different classes: strongly supervised and weakly supervised

methods. In the case of strongly supervised methods, only class labels are available for individual modalities. However, for weakly supervised methods image-text pairs are presented during training. Normally it is assumed that training and testing classes are pre-defined in strongly supervised methods. However, this is problematic for many reasons. First, it is highly dependent on labeled data, which may be unavailable due to annotation costs. Second, annotation requires highly qualified image experts. Third, in practice most of the real-world problems in searches fall into the category of weakly supervised methods i.e. in the social media domain. In this paper, therefore, we focus on weakly supervised approaches that have prevailing fundamental research and application importance in realistic scenarios.

This paper presents a novel resource evaluation dataset for cross-media searching, called FB5K along with a benchmark

learning system. Existing cross-media or multi-modal retrieval datasets have some limitations. Firstly, some datasets they have shortcomings in media types and categories such as, the Wikipedia dataset<sup>1</sup> [8], which contains only two types of media (images and text). Similarly, the Pascal VOC 2012 dataset<sup>2</sup> [9] consists of only 20 categories. However, cross-media retrieval implicates numerous domains under real-world internet conditions. Cross-media retrieval systems trained on scanty domain datasets have difficulties in handling queries from anonymous domains. Secondly, some datasets lack context information i.e. link relations. Such context information is quite accurate, and can provide significant evidence to ameliorate cross-media retrieval system accuracy. Thirdly, popular cross-media datasets are small in size, for example Xmedia [10], IAPR TC-12 [11], and Wikipedia. This deficiency in appropriate data makes it difficult for retrieval systems to learn and evaluate the robustness in real-world galleries. Fourthly, datasets such as, ALIPR [12], SML [13], either just used all the image annotation keywords associated with training images, or unenforced any constraint to the annotation vocabulary for example ESP [14], LabelMe [15], and AnnoSearch [16]. Therefore, these datasets essentially neglect the differences among keywords relating to semantic gaps.

Despite their above-mentioned limitations, these efforts nonetheless provide a significant contribution to the cross-media research community in terms of concept corpus setting, and thus open the gateway for researchers to emphasize ongoing-work on a clear set of semantics. Nevertheless, we suggest that, realistically, semantic gaps are non-uniform in a low-level feature space, and that neglecting such semantic gap differences is inappropriate. For instance, modeling a broad theme, for example *Asia*, is more challenging than modeling a specific theme, for example *sky*, due to the absence of a significant, unique visual feature that can characterize the concept of *Asia*. In addition, most of the time, we typically select local or color features to model concepts like *sky* or *sunset*, respectively.

Considering the aforementioned problems, this paper makes three major contributions. The first is the collection of a new resource evaluation cross-media retrieval dataset, named FB5K. It contains 5130 image-feeling pairs collected from Facebook,<sup>3</sup> introduced for the first time in the cross-media retrieval research community. This dataset is differentiated from current datasets in three aspects: varied domains, high-level semantic information incorporation, and rich context information. Eventually, it should provide a more accurate standard for cross-media study. Therefore, we constructed a standard dataset, keeping in mind the research issues to focus researcher/developer efforts on cross-media retrieval algorithm development, instead of laboriously comparing methods and results. The second is that, to the best of

our knowledge, this is the first effort to collect a dataset of high-level concepts with small semantic gaps based on users' semantic descriptions i.e. image-feeling relationships. Third, this approach aims to learn the cross-media embeddings of users' feelings, images, and tags/texts. We propose a novel method by using Optical Character Recognition (OCR), explicit incorporation of high-level semantic information, and a new similarity measurement in the embedded space, which significantly overcomes the conventional Euclidean distance and improve retrieval performance.

The organization of the rest of this paper is as follows: Section II describes related work. We then describe the characteristics, collection, potential applications, and some example images from the proposed dataset in section III. In Section IV we propose a method to incorporate OCR, high-level semantic information and a specially developed similarity measurement in the embedded space, into existing retrieval methods. The standard methods, evaluation matrices and experimental results are discussed in Section V. Finally, we conclude our paper and provide some useful future directions in section VI.

## II. RELATED WORK

This section describes the related work on cross-media retrieval approaches.

### A. RETRIEVAL METHODS IN CROSS-MEDIA

#### 1) COMMON SPACE LEARNING METHODS

These are the most typical learning methods used in current cross-media retrieval systems. They learn a common space for cross-media data, originated from single model subspace approaches [17]. These methods explicitly project different modality data to a common space for similarity measurement.

There are seven different categories based on common space learning methods. A detailed explanation of individual methods is beyond the scope of this paper; however, here we summarize each method.

- 1) *Statistical correlation analysis methods* [18]–[20]. These are the fundamental models that provided the groundwork for common space learning methods, which optimize statistical values by learning the linear projection matrices for mutual spaces.
- 2) *DNN-based methods* [21]–[27]. Deep Neural Networks (DNNs) are used as fundamental models in cross-media retrieval methods due to their robust abstraction capabilities.
- 3) *Cross-media graph regularization methods* [28]–[30]. These approaches characterize complex cross-media correlations through graph models.
- 4) *Metric learning methods* [31]. These methods view cross-media correlations as a set of similar/different constraints.
- 5) *Learning to rank methods* [32], [33]. These methods emphasize cross-media ranking statistics as their optimization objective.

<sup>1</sup><http://www.svcl.ucsd.edu/projects/crossmodal/>

<sup>2</sup><http://host.robots.ox.ac.uk/pascal/VOC/>

<sup>3</sup>[facebook.com](https://www.facebook.com)

TABLE 1. A summary of multi-modal datasets.

Dataset	Modality	No. of samples	Image features	Text feature	Categories
Wikipedia	image-text	2,866	SIFT+ BOW	LDA	10
NUS-WIDE	image/tags	186,577	6 types	tag occurrence feature	81
Pascal-VOC	image/tags	9,963	3 types	tag occurrence feature	20
Flickr30K	image/sentences	31,783	-	-	-
Twitter-100K [47]	Image-text	100,000	-	-	-
INRIA-Websearch	Image-text	71,478	-	-	353

- 6) *Dictionary learning methods* [34], [35]. These methods produce vocabularies and a learned common space for sparse coefficients of cross-media datasets.
- 7) *Cross-media hashing methods* [23], [36], [37]. The main objective of these methods is to accelerate retrieval performance by learning a common hamming space.

## 2) SIMILARITY MEASUREMENT METHODS

These methods are used to directly measure similarities between different modalities, without explicitly projecting media instances from different modalities spaces to a mutual space. Distance measurements or other typical classifiers cannot be used directly for computing cross-media similarity measurements in the absence of a common space. An instinctive way of linking the “media gap” is by using the known media instances usage and correlations in datasets as the origin. For most of the current methods [38], [39], cross-media similarity measurements represent the relationships between media instances and multimedia documents (MMDs) by using edges in graphs. Cross-media similarity measurement methods can be further categorized into two different approaches:

- 1) *Graph-based methods* [40], [41]. These methods concentrate on the construction of graphs.
- 2) *Neighbor analysis methods* [42], [43]. For these methods local relationships between data are considered for similarity measurement.

## 3) OTHER METHODS

In addition to the aforementioned methods, we further categorized cross-media retrieval into two different sub-classes of other methods:

- 1) *Relevant feedback analysis* [40], [44]. This is a useful method for bridging a huge “media gap” by providing additional information on user intention to facilitate cross-media retrieval performance.
- 2) *Multimodal topic models* [45], [46]. These models are used extensively in cross-media applications. They outlooks cross-media data at the topic level and frequently obtain cross-media similarities by calculating conditional probabilities.

## B. POPULAR CROSS-MEDIA DATASETS

Datasets play a key role in the assessment of cross-media retrieval methods. Table 1 depicts a summarized evaluation of some well-known cross-media datasets.

### 1) XMEDIA DATASET<sup>4</sup>

The only cross-media dataset with five different modalities: video, audio, image, text and 3-Dimensional (3D) model. It contains a further 20 different categories, for example bird, dog, explosion, elephant etc. For individual categories, it has 600 media instances: 250 texts, 250 images, 25 videos, 50 audio clips, and 25 3D models. Hence, the overall number of media instances is 12000. For this dataset, all data is extracted from well-known websites such as Flickr, YouTube, Wikipedia, 3D Warehouse, and Princeton 3D model search engine, as shown in Fig. 1 row 1.

### 2) FLICKR30K DATASET<sup>5</sup> [48]

This dataset is the extension of a previously published dataset called Flickr8K [49], which contains 31783 images collected from the social media website Flickr. Individual images are independently linked with five native English speakers’ descriptive sentences. It lacks category information but has a high correlation between images and text, which makes it a challenging dataset for cross-media retrieval. Nevertheless, Flickr30K lacks content diversity as it focuses only on people’s involvement in daily activities, occasions, and scenes. Fig. 1, row 2, shows examples of this dataset.

### 3) WIKIPEDIA DATASET

This is the most frequently used dataset for the performance measurement of learning systems in cross-media retrieval. It contains 2866 image-text pairs from 10 different classes collected from Wikipedia’s featured articles. However, it is small-scale and contains only two modalities (image and text). Moreover, the classes are of high-level semantics that are hard to distinguish, for example war and history, and thus most of the words provide no interpretations of the visual images. Fig. 1, row 3, shows some examples from this dataset.

### 4) THE PASCAL VOC DATASET<sup>6</sup> [9]

This dataset contains 20 different classes with 5011 training and 4952 testing image-tag pairs. As some images are multi-labeled, previous studies have selected those images which have only one object, resulting in 2808 training and 2841 testing pairs [50]. GIST and color [9], histograms of bag-of-visual-words are the image features whereas; 399-dimensional tag occurrence features are the text features. Some examples are shown in Fig. 1, row 4.

<sup>4</sup><http://www.icst.pku.edu.cn/mipl/XMedia>

<sup>5</sup><http://shannon.cs.illinois.edu/DenotationGraph/>

<sup>6</sup><http://www.cs.utexas.edu/~grauman/research/datasets.html>





## B. DATASET CHARACTERISTICS

The performance of cross-media retrieval methods is highly dependent on the nature of the dataset used for their evaluation. The FB5K dataset includes a set of images that are closely associated with user feelings. These images were crawled from Facebook along with the user-associated feelings. The FB5K dataset has the following attributes:

- First, since this dataset was collected from a social media website, it contains a broad variety of domains under single examples of feelings such as, *hungry*, *love*, *sad*, *thankful* etc.
- Second, the relationship between images and users' feelings is often very strong. In the examples given in Fig. 2, the images have strong ties with the associated feelings. Such is the case in a realistic scenario.
- Third, FB5K is a large-scale dataset, containing 5130 image-text pairs, which helps to avoid overfitting during system training. In other words, it helps to test the cross-media retrieval method's robustness via a wealth of data.
- Fourth, this dataset helps to reduce the semantic gap by providing more accessible visual content descriptors using high-level semantic concepts.

To our knowledge, this is the first cross-media dataset that consists of the above-mentioned characteristics. Also, we believe that FB5K is the first dataset collected from Facebook that comprises high-level concepts with minor semantic gaps between users' semantic descriptions, and a ground-truth of 70 concepts for the whole dataset.



FIGURE 2. Feeling happy examples (smiling, laughing).

## C. POTENTIAL APPLICATION SCENARIO

FB5K provides a more practical standard for cross-media retrieval. The potential application scenarios are outlined as follows:

- A social media website, i.e. Facebook provides predefined emoticons for users to pick at the time of posting a tweet. These emoticons are highly correlated with the posted image and the user's interpretation of the image. Hence, it is more useful and interesting to link the range of emoticons with a user image and recommend a suitable image according to his/her feelings about the contents of the post.
- Social network use has produced a huge amount of multimedia data on the internet, which remains poorly organized, while annotations are time-consuming and expensive. Labeling such large-scale multi-modal data is challenging. Adding user feelings to images can



FIGURE 3. Feeling sad examples (crying, serious, lose).

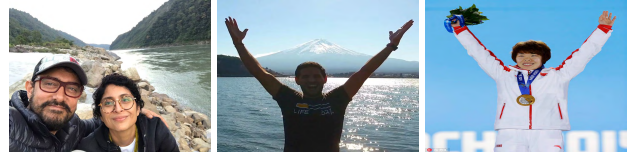


FIGURE 4. Feeling wonderful examples (mountain, river, medal).



FIGURE 5. Feeling cold examples (cap, shivering, jacket, snow).



FIGURE 6. Feeling hungry examples (food, person).



FIGURE 7. Feeling love examples (kissing, hugging).



FIGURE 8. Feeling excited examples (traveling, luggage, vehicle, road).

improve the learning rate of semantic correlations among multi-modal data.

## D. EXAMPLE IMAGES

The dataset includes a wide range of images with associated user feelings, for example *happy* (Fig. 2), *sad* (Fig. 3), *wonderful* (Fig. 4), *cold* (Fig. 5), *hungry* (Fig. 6), *love* (Fig. 7), *excited* (Fig. 8) and *thankful* (Fig. 9).

## E. DIVERSITY OF THE IMAGE COLLECTION

The FB5K dataset comprises numerous images of similar poses but varying illumination level, viewing angles and backgrounds. The reason is that most images uploaded on



FIGURE 9. Feeling thankful examples (wedding, birthday, cake, dancing).



FIGURE 10. Same user feeling from different viewing angles.



FIGURE 11. Same user feeling with different background.



FIGURE 12. User feelings captured at a different time of day, for example in the morning, during the day, and at night.

social websites with the same associated feelings have similar visual content. For example, *smile* is common among different people stating a feeling of happiness, as shown in Fig. 2. However, the viewing angle (Fig. 10), background (Fig. 11) and time of day (Fig. 12) varies. This makes the standard FB5K dataset suitable for content-based retrieval tasks, as it permits a variety of exemplary quests to explore the efficiency of retrieval systems with these fluctuating settings.

#### IV. PROPOSED RETRIEVAL METHOD FOR THE FB5K DATASET

This section briefly explains the proposed cross-media retrieval algorithm for FB5K. Numerous features are used for image representation, for example SIFT [52], color features [10], [53], GIST [54] and HOG [55], [56]. These features are useful for extracting colors and shapes of images, but not for words represented by the images. In this regards, we first propose OCR then adopt explicit incorporation of high-level semantic information and finally develop a novel similarity measurement in the embedded space to improve the retrieval performance.

##### A. IMPLEMENTATION

We summarized our model for cross-media retrieval in algorithm 1. However, a detailed explanation is provided as follows:

#### Algorithm 1 Cross-Media Retrieval Algorithm for FB5k

- 1: **Initialization**
- 2: Word extraction using tessart;
- 3: Vocabulary generation of the most recurrent 2,000 words using the tweets in FB5K;
- 4: **end initialization**

##### Stage 1: High-level semantic information incorporation

- 5:  $K_x(i, j) = \psi_x(i)\psi_x(j)^T$ ;
- 6:

$$\min_{w_1, w_2, w_3} = \sum_{x, y=1}^3 \|\psi_1(I)W_1 - \psi_2(T)W_2\|_2 + \|\psi_1(I)W_1 - \psi_3(C)W_3\|_2 + \|\psi_2(T)W_2 - \psi_3(C)W_3\|_2$$

##### Stage 2: Similarity or distance measurement

- 7:

$$\text{sim}(x_i y_j) = \frac{(\psi_x(i)W_x)(\psi_y(j)W_y)^T}{\|\psi_x(i)W_x\|_2 \|\psi_y(j)W_y\|_2}$$

#### 1) TEXT EXTRACTION

First is the extraction of words on each image using tessart.<sup>9</sup>

#### 2) INCORPORATION OF HIGH-LEVEL SEMANTIC INFORMATION

To facilitate OCR text extraction we incorporate high-level semantic information for learning a common space for image, text/tag, and semantic information (user feelings). Assume we have  $n$  training images having  $i_f$ -dimensional visual feature vectors and  $t_f$ -dimensional tag feature vectors. where  $I \in \mathbb{R}^{n \times i_f}$  and  $T \in \mathbb{R}^{n \times t_f}$ . Furthermore, we also associated each training image with a high-level semantic class,  $C \in \mathbb{R}^{n \times c}$ , where  $c$  represents the number of categories. Individual images are labeled with one  $c$  class (only one specific class in each row of  $K$  is 1 and the remaining are 0).

Let  $i, j$  denote two points. We define similarity as:

$$K_x(i, j) = \psi_x(i)\psi_x(j)^T \quad (1)$$

Where  $K_x$  is a kernel function and  $\psi_x(\cdot)$  represent a function embedding the original feature vector into a nonlinear kernel space.

The goal is to find matrices  $W_x$  that project the embedded vector  $\psi_x(i)$  to minimize the distance between data items.

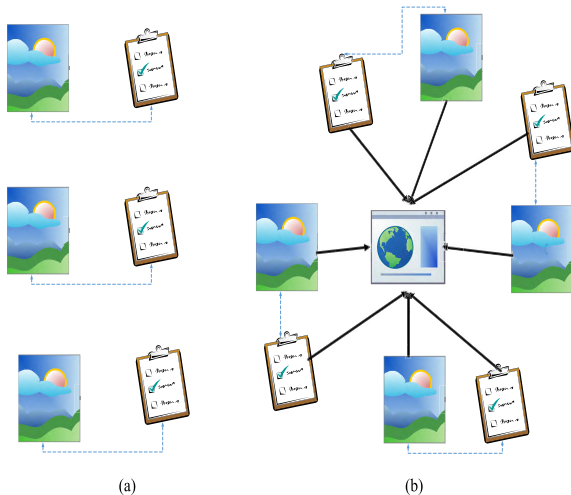
The objective function can be mathematically expressed as:

$$\begin{aligned} \min_{w_1, w_2, w_3} &= \sum_{x, y=1}^3 \|\psi_1(I)W_1 - \psi_2(T)W_2\|_2 \\ &+ \|\psi_1(I)W_1 - \psi_3(C)W_3\|_2 \\ &+ \|\psi_2(T)W_2 - \psi_3(C)W_3\|_2 \\ &\text{where } w_1 = w_2 = w_3 = 0 \end{aligned} \quad (2)$$

<sup>9</sup><https://github.com/tesseract-ocr/tesseract>



This equation tries to align corresponding images and tags [57], whereas, the remaining two terms try to align images with their semantic class. Fig. 13 illustrates graphically the benefits of incorporating high-level semantic information.



**FIGURE 13.** Graphical representation of high-level semantic information incorporation: (a) without semantic class and (b) with semantic class.

### 3) SIMILARITY MEASURE

The similarity measure is an important function used to measure the similarity between text and image. An obvious choice is the Euclidean or Jaccard distance among embedded data points as used in [58] and [47], respectively. However, for our learned embedding, we developed a novel similarity measurement that yielded better realistic results. Mathematically, this can be expressed as:

$$\text{sim}(x_i y_i) = \frac{(\psi_x(i)W_x)(\psi_y(j)W_y)^T}{\|(\psi_x(i)W_x)\|_2 \|(\psi_y(j)W_y)\|_2} \quad (3)$$

Where  $x_i$  represents the training image and  $y_i$  represents the corresponding tweet.  $W_x$  projects the embedded vector  $\psi_x(i)$  and  $W_y$  projects the embedded vector  $\psi_y(j)$  to minimize the distance between image and text.

### 4) DISTANCE IN COMMON SUBSPACE

In this paper we represent the cosine distance between two different modalities in the common subspace as  $\text{Cos}(Twt, Img)$ , where  $Twt$  and  $Img$  represent the tweet and image. It was learned by retrieval methods such as Corr-AE and subspace methods.

### 5) RANKING

Each candidate in the gallery was ranked, based on similarity distances between the queries and candidates.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

We did several experiments to assess the retrieval performance. It is clear from the experimental results that four components affected the final results – dataset repre-

sentation, distance or similarity measures, text features and retrieval methods.

### A. EXPERIMENTAL SETUP

All experiments were performed on four subspace learning methods, which were Canonical Correlation Analysis (CCA) [8], Bilinear Model (BLM) [59], Partial Least Square (PLS) [60], and Generalized Multi-view Marginal Fisher Analysis (GMMFA) [50] and three Corr-AE methods [61]: Corr-AE, cross Corr-AE and full Corr-AE.

In the case of subspace learning methods, we used the implementation from [50] to compute the linear projection matrix. For Corr-AE methods, we use the implementation of [61] to calculate the hidden vectors of the two different modalities. We employed a 1024-dimensional hidden layer. For Corr-AE, cross Corr-AE and full Corr-AE the weight factors for reconstruction errors and correlation distances were set to 0.8, 0.2 and 0.8, respectively.

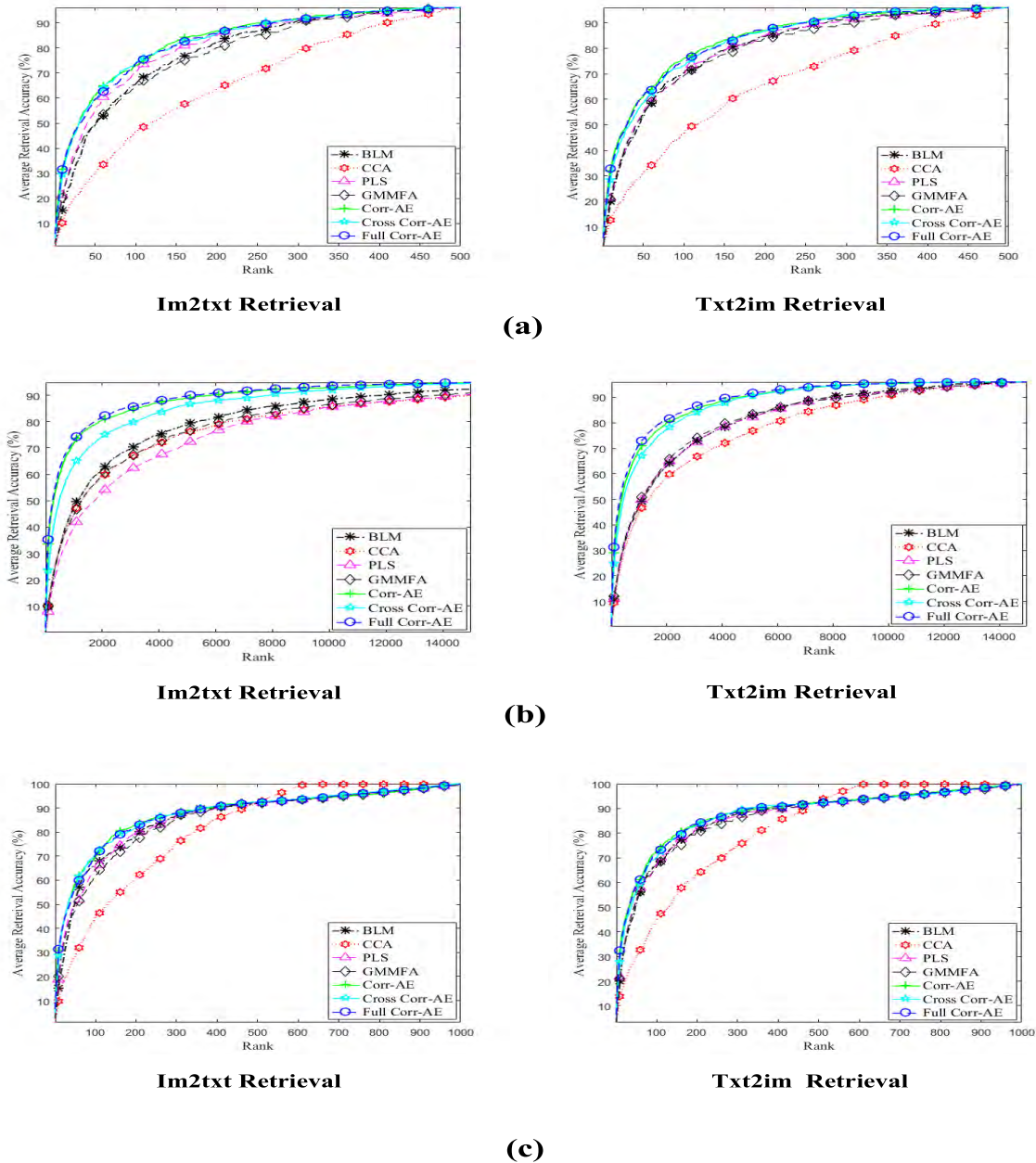
### 1) DATASET SPLITTING

We used three datasets in each experiment: Wikipedia, Flickr30K and FB5K. We split each dataset into a training set, a testing set, and a validation set, as illustrated below:

- 1) *Wikipedia dataset.* In the case of subspace learning, we used 2173 and 500 image-text pairs for training and testing respectively, while for Corr-AE methods a further 193 pairs served as a validation set. We utilized all of the data in a test set as a query.
- 2) *Flickr30k dataset.* For subspace learning, we used 15000 image-text pairs for both training and testing while for Corr-AE methods an additional 1783 image-text pairs were added for validation. We randomly selected 2000 images and texts from the test set to function as a query. As each image had five associated sentences, when taking images as queries, the text gallery comprised 75000 sentences, with any one of the five associated sentences considered to be correct.
- 3) *FB5K dataset.* We split the dataset into 80% and 20% image-text pairs for training and testing respectively. We used the same split for subspace learning, while for Corr-AE methods 250 additional image-text pairs served as a validation set.

### 2) REPRESENTATION

All images were first resized to dimension of  $224 \times 224$ . Then we extracted the last fully connected (fc7) Convolution Neural Network (CNN) features using VGG16 [18] with CAFFE [62] implementation. Text representation was based on Latent Dirichlet Allocation (LDA) [63]. An LDA model was learned from all texts and used to compute the probability of each text under 50 hidden topics. We used this probability vector for text representation. A Bag-of-Word (BoW) model was used for text representation in Corr-AE methods. Initially, texts were converted into lower case, with



**FIGURE 14.** CMC curves compared for different Corr-AE and subspace learning methods using different cross-media datasets: (a) Wikipedia, (b) Flickr30k, and (c) FB5K.

all stopping-words removed. A unigram model was adopted to form a dictionary of the most recurrent 5000 words. Based on this dictionary, for each text we generated a 5000-dimensional BoW model.

### 3) EVALUATION PARAMETERS

We assessed the retrieval performance using Cumulative Match Characteristic (CMC) curves and mean rank. CMC is a useful approach that is used as the evaluation metric in many applications such as face recognition [64], [65] and biometric systems [66]–[68]. For cross-media retrieval, CMC can be illustrated by a curve of average retrieval accuracy with respect to the average ranks of the correct matches for a series

of queries,  $K$ , where rank is:

$$Rank = \frac{1}{|K|} \sum_{x=1}^K rank_x, \tag{4}$$

$rank_x$  refers to the rank position of the correct match for the  $x^{th}$  query.

### B. RETRIEVAL METHODS COMPARISON USING DIFFERENT DATASETS

We tested different cross-media retrieval methods on Flickr30k, Wikipedia and FB5K datasets. Fig. 14 shows the effectiveness of the different retrieval methods. We drew several conclusions from this.



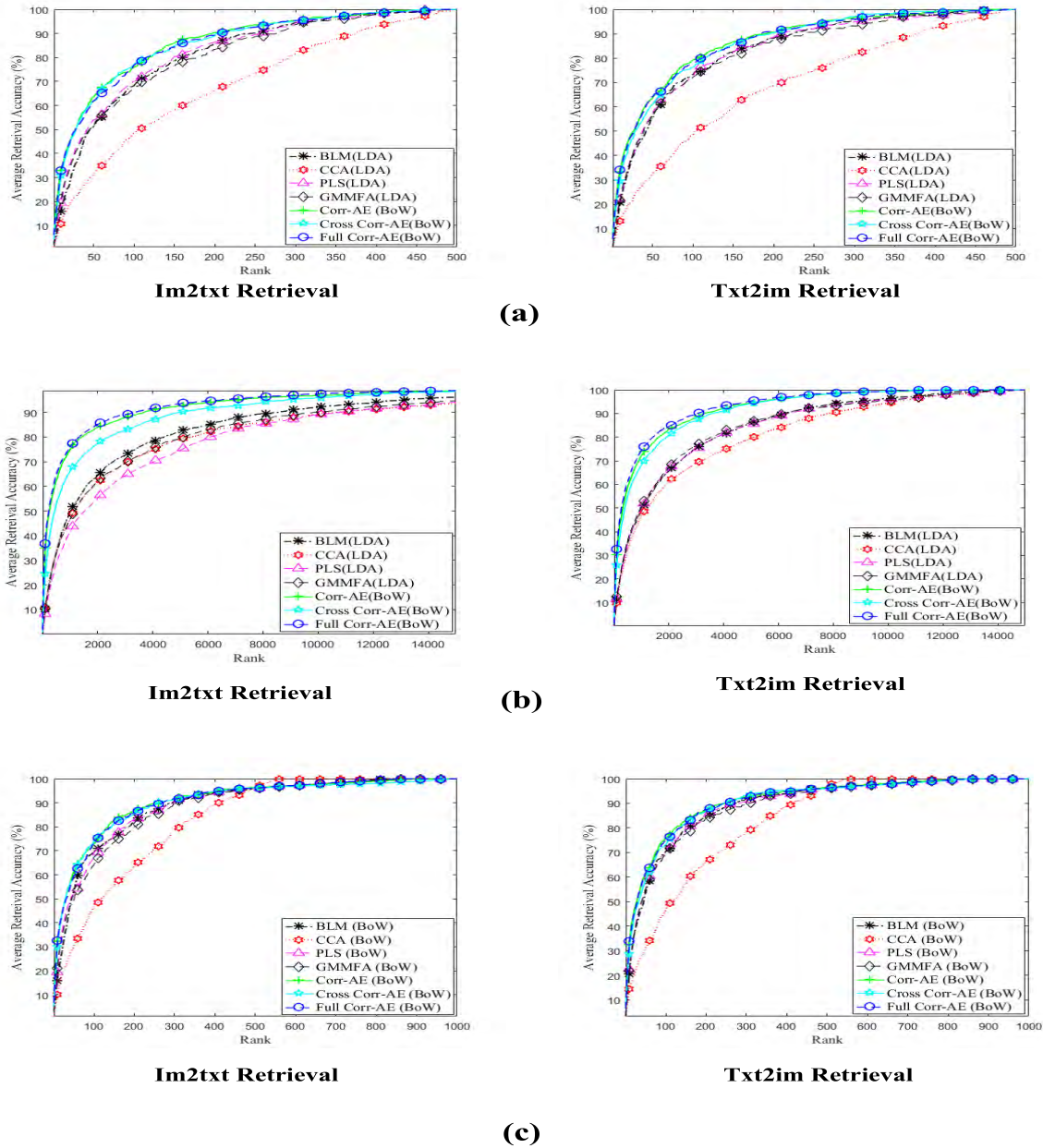


FIGURE 15. CMC curves with different text features for the (a) Wikipedia, (b) Flickr30k, and (c) FB5K datasets.

Corr-AE methods performed well compared to subspace learning methods with all three datasets. However, CCA showed a significant improvement in performance as the number of training samples increased using FB5K. The logic behind this is that correlation is ignored between different modalities in subspace learning when representation learning is performed. However, representation learning and correlation learning are merged into a single process in Corr-AE methods. Furthermore, Corr-AE is used to train a model by minimizing linear combinations of representation learning error and correlation learning error for individual modalities, and between hidden representations of two modalities [56].

This minimization of correlation learning error helps the model in learning hidden representations, while minimization of representation learning error makes better hidden representations to reconstruct the input of individual modalities.

The retrieval performance was highest for FB5K and lowest for Wikipedia. This shows that the tweets are highly correlated when using FB5K compared with Flickr30k and Wikipedia. The reason that FB5K obtained the highest retrieval accuracy is twofold: first, it contained high-level concepts with small semantic gaps. Second, text and images were highly correlated in this dataset. We conclude that user descriptions on tweets are highly correlated to the scenarios.

### C. PERFORMANCE COMPARISON OF DIFFERENT TEXT FEATURES

In this subsection, we assess the performance of cross-modal retrieval learning methods on different text features using Flickr30k, Wikipedia and FB5K. The results are illustrated in Fig. 15.

We applied an LDA text feature on subspace learning methods and a BoW feature on Cross-AE methods. We observed that both of these text features improved the performance of the learning methods. For both Im2txt and Txt2im, consider Figs. 15(a) and (b), it is clear that the exploitation of LDA and BoW features improved average retrieval accuracy by 2.2% for CCA at rank = 2000, and Corr-AE at rank = 1500 for the Flickr30k dataset. The improvement in performance continued for CCA, PLS and Corr-AE, which achieved increases in accuracy of 4.6%, 3.8% and 6.2% at ranks 180, 200 and 250, respectively for the Wikipedia dataset. For FB5K, we used only a BoW feature, a similar performance was observed; full Corr-AE and GMMFA raised the accuracy from 72.5% to 82.5% and 67.2% to 80.5% respectively, at rank = 200.

This increase in the performance of cross-media retrieval methods was due to semantic information brought by the word vectors. Moreover, for all three datasets the images were well explained in formal language or, in other words, the images and text/tags were highly correlated. Therefore, both of the text features (LDA and BoW) were productive in text representation and offered an advantage over the baseline retrieval methods for all of the datasets.

### D. PROPOSED METHOD PERFORMANCE

In this section, we describe the proposed method for evaluation of FB5K. We compared its performance with the baseline methods. Fig. 16 shows the experimental results.

Fig. 16 clearly shows that using the proposed method in the baseline learning systems significantly improved their performance. In particular, using OCR, explicit incorporation of high-level semantic information, and a specially developed similarity measurement in the embedded space improved cross-media retrieval accuracy when similar retrieval methods were used. For example, in the case of Txt2im retrieval, CCA achieved 45% accuracy at rank = 110, whereas the BLM, PLS, and GMMFA methods achieved the same accuracy at ranks 20, 25 and 18, respectively. Incorporating the proposed method boosted the accuracy of CCA, BLM, PLS, and GMMFA to 6.5%, 4%, 5% and 7% respectively, at the same rank.

The proposed retrieval method significantly increased the retrieval accuracy of both the baseline methods, i.e. subspace learning methods and Cross-AE methods, due to the following three facts. First, incorporation of OCR used the text information inside the images along with color and shape information. Second, incorporating high-level semantic information produced a third view-class for data of different modalities (text and image). Third, the proposed similarity

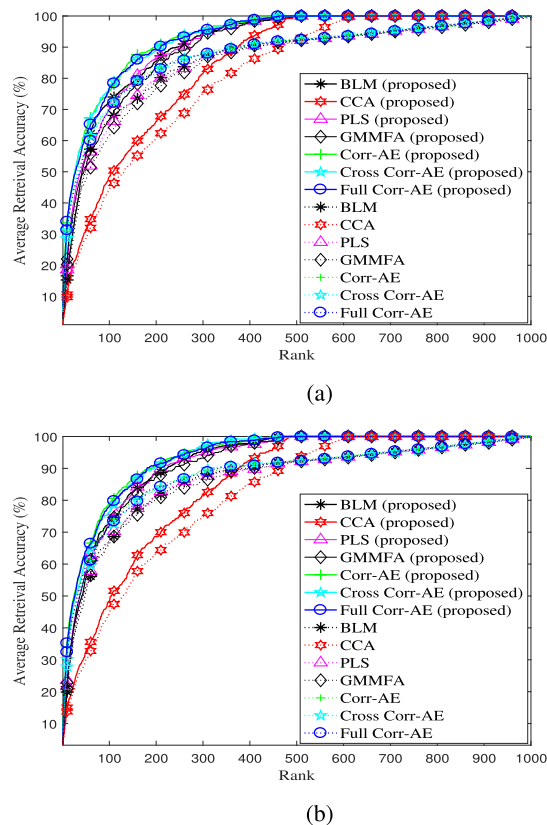


FIGURE 16. CMC curves on FB5K with the proposed method and baselines. (a) Im2txt retrieval. (b) Txt2im retrieval.

measure helped in adjusting the weights to minimize the distances between image and text.













### E. FB5K RETRIEVED EXAMPLES

This section describes the retrieval examples for FB5K using CCA and the proposed method.

Fig. 17(a) shows the retrieval image results for different query tags. It shows that the proposed method was successful in learning colors, background and class information, e.g. in Fig. 17(a) we used the keyword *cold* to retrieve the images on right, which strongly indicate that the keyword information lay in the retrieved image. Moreover, incorporation of semantic class not only improved the retrieval accuracy, but also provided higher weights to more minor concepts during the formation of query tag vectors.

Fig. 17(b) shows the tagging results retrieved by the proposed method on some test images. It is clear that using the proposed method with FB5K significantly outperformed the baseline methods, despite its diverse features.

Furthermore, explicit incorporation of high-level semantic information and a novel developed similarity function in the embedded space improve retrieval performance. For example, in Fig. 17(b), the query image contains snow, women in jacket, mountain and trees. Despite the original class retrieval being *cold* and *shivering*, opinion words (*beautiful*, *travel*) also appeared in the retrieved texts, which show the

Query tag	Retrieved images				
Cold					
Hungry					
(a)					
Query Image	Retrieved tags				
	Happy	Love	Laughing	Person	Shivering
	Cold	Shivering	Beautiful	Travel	Scenery
(b)					

**FIGURE 17.** Retrieval examples for FB5K using CCA and the proposed method. The first two rows represent the query tag and its corresponding top five retrieved images, whereas the last two rows show query images and their corresponding top five retrieved tags. (a) tag/txt2img retrieval. (b) img2tag/txt retrieval.

significance of FB5K as it also covers sentiment and opinion of the users.

FB5K provides information that is more realistic to the user. It incorporates high-level semantic information by providing the class probability for individual images. For example, in Fig. 17(b), with a query image of a baby, the proposed method retrieved *happy* and *love* as high frequency words in the retrieved text. This shows that despite the sentiment of an image being hidden under high-level concepts, opinion characteristics can have an impact on multi-modal retrieval.

## VI. CONCLUSION AND FUTURE WORK

This paper introduced a novel cross-media dataset called FB5K. We also presented a more realistic embedding approach for images, tags/texts, and their semantics. Specifically, in order to learn the cross-modal embeddings of user feelings, images and tags/texts, we developed a novel method by utilizing OCR, explicit incorporation of high-level semantic information, and a new similarity measurement in the embedded space, to improve the retrieval performance.

However, much effort will still be needed in the future, since no perfect system has yet been developed. In future work, we plan to increase the number of images and user feelings in the FB5K dataset and develop a robust and effective evaluation protocol. We believe that FB5K and the proposed cross-media retrieval method suffice as a reference guide

for researchers and developers to facilitate the design and implementation of better evaluation protocols.

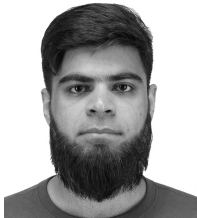
## REFERENCES

- [1] Y.-J. Lu, P. A. Nguyen, H. Zhang, and C.-W. Ngo, "Concept-based interactive search system," in *Proc. Int. Conf. Multimedia Modeling*. Springer, 2017, pp. 463–468.
- [2] R. A. Kambau and Z. A. Hasibuan, "Concept-based multimedia information retrieval system using ontology search in cultural heritage," in *Proc. IEEE 2nd Int. Conf. Informat. Comput. (ICIC)*, Nov. 2017, pp. 1–6.
- [3] Z. Ma, H. Yu, W. Chen, and G. J., "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, 2018.
- [4] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.
- [5] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.
- [6] Z. Ma et al., "The role of data analysis in the development of intelligent energy networks," *IEEE Netw.*, vol. 31, no. 5, pp. 88–95, 2017.
- [7] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognit.*, vol. 47, no. 9, pp. 3143–3157, 2014.
- [8] N. Rasiwasia et al., "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.
- [9] S. J. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1145–1158, Jun. 2011.
- [10] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.



- [11] M. Grubinger, P. Clough, and H. Muller, and T. Deselaers, "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *Proc. Int. Workshop OntoImage*, vol. 5, 2006, p. 10.
- [12] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 985–1002, Jun. 2008.
- [13] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 394–410, Mar. 2007.
- [14] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2004, pp. 319–326.
- [15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.
- [16] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma, "AnnoSearch: Image auto-annotation by search," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1483–1490.
- [17] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, and C. Zhang, "Compound rank- $k$  projections for bilinear analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1502–1513, Jul. 2016.
- [18] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Artificial Intelligence and Statistics*. 2014, pp. 823–831.
- [19] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4094–4102.
- [20] J. C. Pereira et al., "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 521–535, Mar. 2014.
- [21] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proc. IJCAI*, 2016, pp. 3846–3853.
- [22] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3441–3450.
- [23] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.
- [24] F. Wu et al., "Learning of multimodal representations with random walks on the click graph," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 630–642, Feb. 2016.
- [25] Y. Wei et al., "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
- [26] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.
- [27] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. (2016). "Learning aligned cross-modal representations from weakly aligned data." [Online]. Available: <https://arxiv.org/abs/1607.07295>
- [28] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.
- [29] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.
- [30] J. Liang, Z. Li, D. Cao, R. He, and J. Wang, "Self-paced cross-modal subspace matching," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2016, pp. 569–578.
- [31] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. AAAI*, 2013, pp. 1–7.
- [32] X. Jiang et al., "Deep compositional cross-modal learning to rank via local-global alignment," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 69–78.
- [33] F. Wu et al., "Cross-modal learning to rank via latent joint representation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1497–1509, May 2015.
- [34] F. Zhu, L. Shao, and M. Yu, "Cross-modality submodular dictionary learning for information retrieval," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, 2014, pp. 1479–1488.
- [35] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI*, 2013, pp. 1070–1076.
- [36] M. Long, Y. Cao, J. Wang, and P. S. Yu, "Composite correlation quantization for efficient multimodal retrieval," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2016, pp. 579–588.
- [37] H. Liu, R. Ji, Y. Wu, and G. Hua. (2016). "Supervised matrix factorization for cross-modality hashing." [Online]. Available: <https://arxiv.org/abs/1603.05572>
- [38] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2016.
- [39] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. neural Netw. Learn. Syst.*, to be published.
- [40] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.
- [41] Y. Yang, F. Wu, D. Xu, Y. Zhuang, and L.-T. Chia, "Cross-media retrieval using query dependent search methods," *Pattern Recognit.*, vol. 43, no. 8, pp. 2927–2936, 2010.
- [42] D. Ma, X. Zhai, and Y. Peng, "Cross-media retrieval by cluster-based correlation analysis," in *Proc. 20th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 3986–3990.
- [43] X. Zhai, Y. Peng, and J. Xiao, "Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval," in *Proc. Int. Conf. Multimedia Modeling*. Springer, 2012, pp. 312–322.
- [44] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 437–446, Apr. 2008.
- [45] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, "Multi-modal mutual topic reinforce modeling for cross-media retrieval," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 307–316.
- [46] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 2407–2414.
- [47] Y. Hu, L. Zheng, Y. Yang, and Y. Huang, "Twitter100k: A real-world dataset for weakly supervised cross-media retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 927–938, Apr. 2018.
- [48] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.
- [49] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, no. 1, pp. 853–899, 2013.
- [50] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multi-view analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2160–2167.
- [51] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world Web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, p. 48.
- [52] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [53] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1939–1946.
- [54] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
- [55] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, vol. 1, no. 1, pp. 886–893.
- [56] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.
- [57] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
- [58] D. P. Foster, S. M. Kakade, and T. Zhang, "Multi-view dimensionality reduction via canonical correlation analysis," *Tech. Rep.*, 2008.
- [59] J. B. Tenenbaum and W. T. Freeman, "Separating style and content," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 662–668.

- [60] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Subspace, Latent Structure and Feature Selection*. Springer, 2006, pp. 34–51.
- [61] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.
- [62] Y. Jia et al., "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
- [63] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [64] N. Khamsemanan, C. Nattee, and N. Jianwattanapaisarn, "Human identification from freestyle walks using posture-based gait feature," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 119–128, Jan. 2018.
- [65] S. U. Rehman, S. Tu, Y. Huang, and Z. Yang, "Face recognition: A novel un-supervised convolutional neural network method," in *Proc. IEEE Int. Conf. Online Anal. Comput. Sci. (ICOACS)*, May 2016, pp. 139–144.
- [66] B. DeCann and A. Ross, "Relating ROC and CMC curves via the biometric menagerie," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep./Oct. 2013, pp. 1–8.
- [67] N. Damer, A. Opel, and A. Nouak, "CMC curve properties and biometric source weighting in multi-biometric score-level fusion," in *Proc. IEEE 17th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2014, pp. 1–6.
- [68] S. N. A. Seha and D. Hatzinakos, "Human recognition using transient auditory evoked potentials: A preliminary study," *IET Biometrics*, vol. 7, no. 3, pp. 242–250, May 2018.



**SADAQAT UR REHMAN** received the B.Sc. degree from the Department of Computer Systems Engineering, University of Engineering and Technology Peshawar, and the M.Sc. degrees from the Department of Electrical Engineering, Sarhad University of Science and Information Technology, in 2011 and 2014, respectively. He has also served at the Sarhad University of Science and Information Technology, Peshawar, Pakistan, as a Lecturer, from 2012 to 2015.

He is currently pursuing the Ph.D. degree with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing China. His current research interests are in the areas of deep learning, including multimedia information retrieval, convolution neural networks, unsupervised learning algorithms, and optimization techniques. He has a range of publications in these fields in the conferences and journals of repute.



**SHANSHAN TU** received the Ph.D. degree from the Computer Science Department, Beijing University of Post and Telecommunication, in 2014. He was with the Department of Electronic Engineering, Tsinghua University, as a Post-Doctoral Researcher, from 2014 to 2016. He visited the University of Essex for joint doctoral training from 2013 to 2014. He is currently an Assistant Professor with the Faculty of Information Technology, Beijing University of Technology, China. His research interests are in the areas of cloud computing security and Deep learning.



**YONGFENG HUANG** (M'10–SM'11) is currently a Professor and the Director of the Research Institute of Information Cognition and Intelligent System, Tsinghua University, Beijing, China. Along his career, he has published five books and over 200 research papers on computer network, machine learning, and multimedia communication. His research interests include Cloud Computing, P2P, multimedia network, Deep learning, and data security.



**OBAID UR REHMAN** received the M.Sc. degree in computer engineering from the University of Liverpool, U.K., and the Ph.D. degree in electrical engineering from Zhejiang University. He is currently an Assistant Professor with the Department of Electrical Engineering, Sarhad University of Science and Information Technology. His research interests are optimization techniques, genetic algorithms, and computational electromagnetics.

...