IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An Efficient Approach for Geo-Multimedia Cross-Modal Retrieval

LEI ZHU[1], JUN LONG[1], CHENGYUAN ZHANG[1], WEIREN YU[2,3], XINPAN YUAN[4], LONGZHI SUN[1]

[1]School of Computer Science and Engineering, Central South University,Changsha, 410083, PR China (e-mail: {leizhu,junlong,cyzhang,sunlongzhi}@csu.edu.cn)
[2]School of Engineering and Applied Science, Aston University, Birmingham, U.K. (e-mail: w.yu3@aston.ac.uk)
[3]Nanjing University of Science and Technology, China (e-mail: w.yu3@aston.ac.uk)
[4]School of Computer Science, Hunan University of Technology, Zhuzhou, 412007, PR China (e-mail: xpyuan@hut.edu.cn)

Corresponding author: Chengyuan Zhang (e-mail: cyzhang@csu.edu.cn).

This work was supported in part by the National Natural Science Foundation of China (61702560, 61472450, 61972203), the Key Research Program of Hunan Province(2016JC2018), project (2018JJ3691) of Science and Technology Plan of Hunan Province, and the Research and Innovation Project of Central South University Graduate Students(2018zzts177).

ABSTRACT Due to the rapid development of mobile Internet techniques, such as online social networking and location-based services, massive amount of multimedia data with geographical information is generated and uploaded to the Internet. In this paper, we propose a novel type of cross-modal multimedia retrieval, called geo-multimedia cross-modal retrieval, which aims to find a set of geo-multimedia objects according to geographical distance proximity and semantic concept similarity. Previous studies for cross-modal retrieval and spatial keyword search cannot address this problem effectively because they do not consider multimedia data with geo-tags (geo-multimedia). Firstly, we present the definition of kNN geo-multimedia cross-modal query and introduce relevant concepts such as spatial distance and semantic similarity measurement. As the key notion of this work, cross-modal semantic representation space is formulated at the first time. A novel framework for geo-multimedia cross-modal retrieval is proposed, which includes multi-modal feature extraction, cross-modal semantic space mapping, geo-multimedia spatial index and cross-modal semantic similarity measurement. To bridge the semantic gap between different modalities, we also propose a method named cross-modal semantic matching (CoSMat for shot) which contains two important components, i.e., CorrProj and LogsTran, which aims to build a common semantic representation space for cross-modal semantic similarity measurement. In addition, to implement semantic similarity measurement, we employ deep learning based method to learn multi-modal features that contains more high level semantic information. Moreover, a novel hybrid index, GMR-Tree is carefully designed, which combines signatures of semantic representations and R-Tree. An efficient GMR-Tree based kNN search algorithm called kGMCMS is developed. Comprehensive experimental evaluations on real and synthetic datasets clearly demonstrate that our approach outperforms the-state-of-the-art methods.

INDEX TERMS Cross-Modal Retrieval, Deep Learning, kNN Spatial Search, Geo-Multimedia

I. INTRODUCTION

Due to the rapid popularity of mobile Internet techniques, online social networking and location-based services, massive amount of multimedia data is generated and uploaded to the Internet. For example, as the largest online social networking site, Facebook[1] has 1.15 billion users registered and the total number of images uploaded is 250 billion since its establishment. Twitter[2] has more than 140 million users who post 400 million tweets in the form of text and image all around the world. In China, the active users of Sina Weibo[3] were 376 million on September 2017. They post and share hundreds of thousands of texts, pictures or videos everyday in this platform. For the photo sharing service, more than 3.5 million new photos were uploaded everyday in 2013 to Flickr[4], which is the most popular photo shared web site and it had a total of 87 million registered users. For the video

[1]https://facebook.com/
[2]http://www.twitter.com/
[3]https://weibo.com/
[4]https://www.flickr.com/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

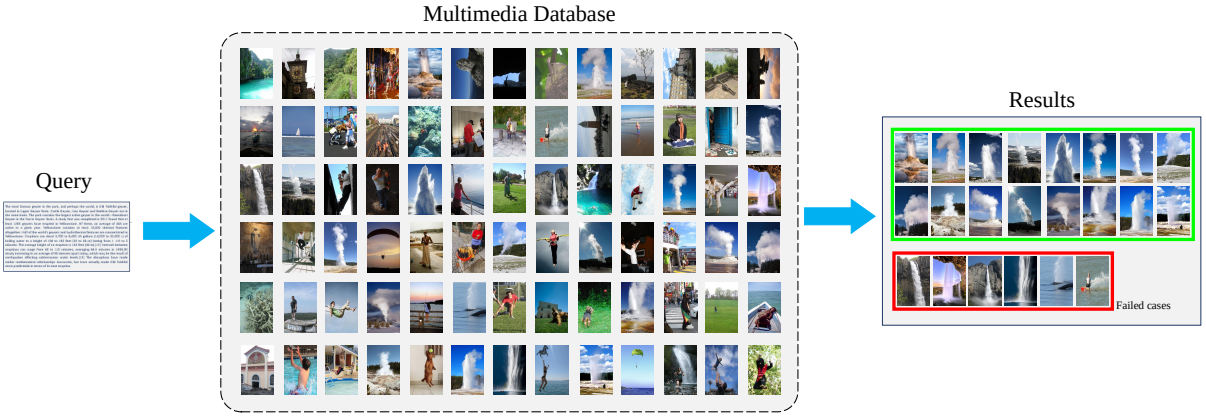


FIGURE 1: An example of cross-modal retrieval. Several images are retrieved from the multimedia database by a textual query. The images in green rectangle are the correct results and the failed cases are in the red rectangle.

sharing service, YouTube[5] shares more than 100 hours of videos every minutes as of the end of 2013. The number of independent users monthly in IQIYI[6], the most popular video website in China, reached 230 million and the total watch time monthly exceeded 42 billion minutes. As the largest online encyclopedia, Wikipedia[7] comprises more than 40 million articles with pictures in 301 different languages. Unlike traditional structured data, these large-scale multimedia [1] data has different modalities [2], e.g. text, image, audio, video. Apparently, the emergence of massive multi-modal data [3], [4] brings great challenges to data storage, mining and retrieval [5]–[7]. This necessitates efficient methods for multimedia data retrieval and processing.

As mentioned above, multi-modal data (text, image, audio, video) describes the world from different perspectives [8]. Each of these modalities corresponds to each perception of human. For instance, our languages can be preserved in the form of text; natural scene can be represented by photos or videos; vocal signals can be recoded in audio files. To ulteriorly imitate human understanding of different modalities and then make search engines have the same capabilities, multi-modal and cross-modal representation and retrieval [9]–[12] problem has been proposed, which involves feature extraction and fusion [13]–[16], representation, semantic understanding, etc. And it is based on many techniques for unimodality retrieval.

Image is one of the most common modalities, and many image retrieval [17] techniques support cross-modal retrieval. Content-based image retrieval (CBIR) is a hot issue in the multimedia area and lots of approaches have been proposed to improve precision and efficiency of image search. Several CBIR systems such as K-DIME [18], IRMFRCAMF [19] and gMRBIR [20] have been proposed to develop advanced multimedia retrieval systems. Moreover, traditional feature extraction methods like scale-invariant feature transform (SIFT) [21], [22] and visual representation model such as bag-of-visual-words (BoVW) [23] are applied in cross-modal retrieval. Recently, CNN [26], [27] based image recognition [24], [25] and retrieval is becoming a hot issue with the rise of deep learning techniques [28]. For instance, [29] reported a quantum jump in image classification, which has the great improvement in performance in ImageNet large scale visual recognition challenge [30]. Other works like [31]–[33] introduced serval new solutions for image search via deep learning.

Another common modality is text, which exists over the Internet environment. Just like image retrieval, text search and understand plays an important role in both natural language processing and information retrieval studies. Many works using deep learning techniques, i.e., CNN [34], LSTM [35], [36], and siamese networks [37] to develop novel solution for semantic textual similarity measurement [38], [39] and retrieval [40].

Unlike the unimodality retrieval above-mentioned, traditional cross-modal retrieval aims to find objects with one modality by the query with another modality. For example, we can issue a query to search an image that can best demonstrate a given sentence or paragraph, or find an article or a poem in text which can describe a given photo. Example 1.1 is an example of traditional cross-modal retrieval.

Example 1.1: Fig. 1 illustrates a typical example of cross-modal retrieval. A user needs to find some pictures about famous geysers. She writes down a short introduction or description of geysers and put it into cross-modal retrieval system. The system then returns several images that are highly relevant to the input text from the multimedia database by cross-modal similarity measurement. Unlike the keyword-based retrieval, cross-modal retrieval is based on understanding of multi-modal data and finding the cross-modal semantic correlation. Clearly, the images in green rectangle are the correct results, which are the photos of geysers. However, the failed cases in the red rectangle are other categories of pictures, i.e., waterfall, spoondrift, water spouts of whales, etc., which are similar to the geysers in the aspect of visual

[5]https://www.youtube.com/
[6]http://www.iqiyi.com/
[7]https://www.wikipedia.org/

content.

As the locating techniques (e.g., GPS and gyroscope) and HD camera are applied widely in smart mobile devices such as smartphones and tablets, massive multimedia data with geo-tags, i.e., geo-images [41], geo-texts and geo-videos have been conveniently collected and uploaded to the Internet. Location-based services such as Google Places and Dianping use geo-texts, geo-images to support spatial object query services, e.g., *Where is the nearest seafood restaurant*, *Which shop nearby sells this type of handbag*. Spatial textual or visual query is a hot spot in the spatial database community, which includes range query [42], kNN query [43], top-k range query [44], interactive query [45], etc. It is concerned by lots of researches these days and several efficient indexing techniques like I^3 [46], KR*-tree [42], IL-Quadtree [47], [48], IR-tree [49] and its variations [50], WIR-tree [51], etc. have been proposed to improve performance of the system.

Motivation. It is a pity that traditional spatial keyword or geo-image queries just consider unimodality during the retrieval. That means these approaches cannot be applied in the cross-modal retrieval directly. On the other hand, previous studies of traditional multi-modal and cross-modal retrieval do not consider the geo-multimedia data. These existing methods cannot improve the retrieval performance by using spatial information. Undoubtedly, geographical location is another significant information for supporting advanced search engines and location-based services. To the best of our knowledge, there is no one who has paid attention on the problem of geo-multimedia cross-modal retrieval at present. To describe this novel retrieval paradigm clearly, a motivating example is introduced below, in which both the cross-modal search and geographical distance proximity are considered.

Example 1.2: As illustrated in Fig. 2, consider a tourist is traveling in a historic city. She is particularly interested in Baroque architecture and wants to visit some ancient buildings in Baroque style. However, she have no idea how many ancient buildings are near her and do not know where these buildings are located. Due to time limit, she cannot seem to go all over the city to find them. In such case, she can write a short paragraph or just a sentence to describe the desirable buildings or the scenery, and put them into search engine as a kNN spatial cross-modal query. The system will return the k nearest ancient buildings geographical location and their photos taken by other people according to her description. With the help of the query, the tourist can find some nearest spots which meet her interests.

In this paper, we aim to combat the challenge described in example 1.2, namely, retrieve a set of results containing k geo-multimedia objects that are nearest to the query location and highly similar to the query in the aspect of semantic concepts. For the first time, we present the definition of a new query paradigm called kNN geo-multimedia cross-modal query and propose a novel score function that consider the geographical distance proximity and semantic similarity between two different geo-multimedia objects. Besides, we introduce the notion of cross-modal semantic representation space and discuss the basic idea of solving cross-modal retrieval. A novel framework of geo-multimedia cross-modal retrieval is presented, which is based on deep learning and spatial indexing techniques. To implement this framework, a novel approach called DeCoSReS is proposed, which employs deep learning techniques to construct a common semantic representation space for different modalities to bridge the semantic gap. In addition, we develop a novel hybrid indexing structure named GMR-Tree that is a combination of signature files and R-Tree to boost the performance. And based on it, an efficient search algorithm named kGMCMS is developed to implement kNN geo-multimedia cross-modal query.

Contributions. The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, this is the first work to investigate the problem of geo-multimedia cross-modal retrieval. We formulate the definition of geo-multimedia object and kNN geo-multimedia cross-modal query, and then propose the notion of cross-modal semantic representation space.
- To solve the problem of geo-multimedia cross-modal retrieval, we introduce a novel framework that consists of multi-modal feature extraction, cross-modal semantic space mapping, geo-multimedia spatial index and cross-modal semantic similarity measurement.
- To bridge the semantic gap between different modalities in the processing of retrieval, we propose a novel approach named CoSMat that consists of two important components i.e., CorrProj and LogsTran. Based on it, a deep learning based method called DeCoSReS is used to generate cross-modal semantic representation.
- To improve the search performance, we present a novel hybrid indexing structure named GMR-Tree which is a combination of signature technique, multi-modal semantic representations and R-Tree. Based on it we develop a novel search algorithm named kGMCMS to boost the retrieval.
- We have conducted extensive experiments on real and synthetic datasets. Experimental results demonstrate that our solution outperforms the-state-of-the-art methods.

Roadmap. The remainder of this paper is organized as follows: the related works are reviewed in Section II. In Section III we introduce the definition of kNN geo-multimedia cross-modal query and relevant concepts. In Section IV, a novel framework of geo-multimedia cross-modal retrieval is proposed. In Section V, we propose the method named cross-modal semantic matching and then a framework of cross-modal semantic representation construction by using deep learning techniques. In Section VI, we design a novel hybrid indexing structure named GMR-Tree and an efficient search algorithm called kGMCMS is developed to support geo-multimedia cross-modal query. Our experimental results are presented in Section VII, and finally we draw the conclusion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

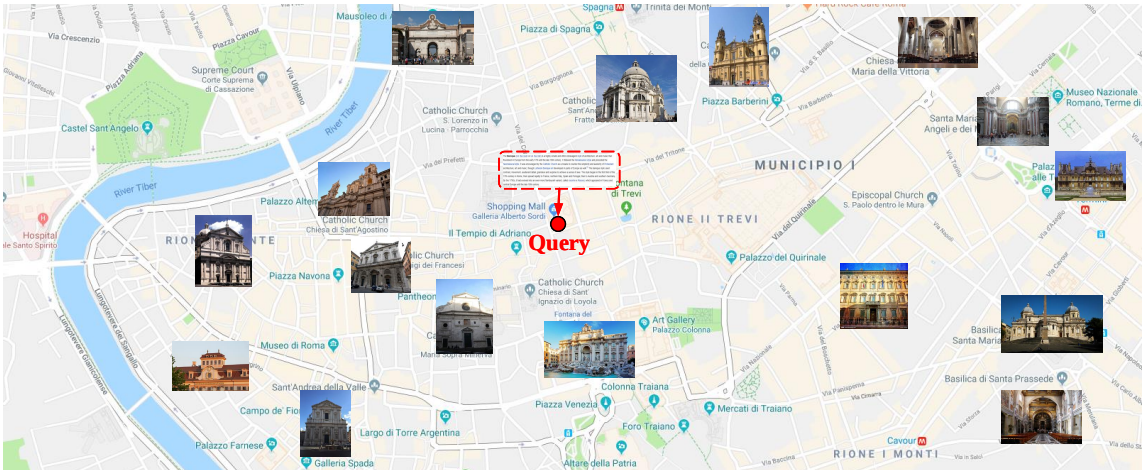


FIGURE 2: An example of kNN spatial cross-modal retrieval.

in Section VIII.

II. RELATED WORK

In this section, we introduce an overview of previous works of multi-modal and cross-modal retrieval, deep learning based multimedia retrieval and spatial textual search, which are related to this work. To the best of our knowledge, there is no existing work on the problem of geo-multimedia cross-modal retrieval.

A. MULTI-MODAL AND CROSS-MODAL RETRIEVAL

Multi-modal and cross-modal retrieval are two hot issues in the field of multimedia analysis and retrieval. A research problem or data set is characterized as multi-modal when it includes multiple modalities [8] such as text, image, audio, video. In the past few years, lots of researchers focus on multi-modal and cross-modal retrieval problem and many significant results have been proposed to improve the retrieval performance.

Multi-Modal Retrieval. Multi-modal retrieval [52] aims to search multimedia data [53] with multiple modalities. Laenen et al. [54] proposed a novel multi-modal fashion search paradigm, which allows users to input a multi-modal query composed of both an image and text. To address this problem, they presented a common, multi-modal space for visual and textual fashion attributes where their inner product measures their semantic similarity. For image raking problem, Yu et al. [55] proposed a novel deep multi-modal distance metric learning method named Deep-MDML to address the two main limitations of similarity estimation in existing CBIR methods: (i) Mahalanobis distance is applied to build a linear distance metric; (ii) these methods are unsuitable for handling multi-modal data [56]. Jin et al. [57] presented a new multi-modal hashing method named SNGH which is to preserve the fine-grained similarity metric based on the semantic graph. They defined a function based on the local similarity in particular to adaptively calculate multi-level similarity by encoding the intra-class and inter-class variations. Rafailidis et al. [58] designed a unified framework for multi-modal content retrieval which supports retrieval for rich media objects as unified sets of different modalities. The main idea is combining all monomodal heterogeneous similarities to a global one according to an automatic weighting scheme to construct a multi-modal space to capture the semantic correlations among multiple modalities. Moon et al. [59] proposed a transfer deep learning (TDL) framework that can transfer the knowledge obtained from a single-modal neural network to a network with a different modality. Several embedding approaches for transferring knowledge between the target and source modalities were proposed by them. Dang-Nguyen et al. [60] proposed a novel framework that can produce a visual description of a tourist attraction by choosing the most diverse pictures from community-contributed datasets to describe the queried location more comprehensively. Based on multi-graph enabled active learning, Wang et al. [61] presented a multi-modal web image retrieval technique to leverage the heterogeneous data on the web to improve retrieval precision. In this solution, three graphes, i.e., Content-Graph, Text-Graph and Link-Graph which are constructed on visual content features, textual annotations and hyperlinks respectively, provide complimentary information on the images. To solve the problem of recipe-oriented image-ingredient correlation learning, Min et al. [62] proposed a multi-modal multitask deep belief network (M^3TDBN) to learn joint image-ingredient representation regularized by different attributes.

Cross-Modal Retrieval. Unlike unimodal retrieval, generally the modalities of query and results are different in cross-modal retrieval, e.g. the retrieval of text documents in response to a query image, and the retrieval of images in response to a query text [63]. To exploit the correlation between multiple modalities, Bredin et al. [64] utilized canonical correlation analysis (CCA) [67] and Co-Inertia Analysis (CoIA) for the task of audio-visual based talking-face biometric verification. Due to the importance of negative correlation, Zhai et al. [65] proposed a novel cross-modality

correlation propagation approach to simultaneously deal with positive correlation and negative correlation between media objects of different modalities. Rasiwasia et al. [66] proposed a novel method named cluster canonical correlation analysis (cluster-CCA) for joint dimensionality reduction of two sets of data points. Based on it they designed a kernel extension named kernel cluster canonical correlation analysis (cluster-KCCA) which achieves superior state of the art performance in cross-modal retrieval task. In another work Rasiwasia et al. [63] studied the problem of joint modeling the text and image components of multimedia documents. They investigated two hypotheses and using canonical correlation analysis to learn the correlations between text and image. To measure the cross-modal similarities, Jia et al. [68] presented a novel Markov random field based model which learns cross-modality similarity from a document corpus that has multinomial data. Chu et al. [69] developed a flexible multimodality graph (MMG) fusion framework to fuse the complex multi-modal data from different media and a topic recovery approach to effectively detect topics from cross-media data.

It is unfortunate that all the researches aforementioned cannot be directly applied to geo-multimedia cross-modal retrieval because they do not consider both the geographical location and multimedia information during the processing of multi-modal or cross-modal retrieval. These solutions are really significant for multimedia information retrieval but they are not adequately suitable to the problem of geo-multimedia cross-modal retrieval. Thus, there is an urgent need to develop efficient methods for geo-multimedia cross-modal retrieval.

B. MULTIMEDIA RETRIEVAL VIA DEEP LEARNING

More recently, lots of multimedia retrieval problems have been solve by new models via deep neural networks [70]–[74]. Content-based image retrieval is one of the significant problems, and many researches improve the retrieval precision with the power of deep learning. Fu et al. [75] proposed a CBIR system based on CNN and SVM. In this framework, CNN is applied to extract the feature representations and SVM is used to learn the similarity measures. A validation set is generated in the training of SVM to tune to parameters. By extending SIFT-based SMK [76], [77]methods, Zhou et al. [78] proposed a unified framework of CNN-based match kernels to encode the two complementary features: low level features and high level features, which can provide complementary information for image retrieval task. To evaluate whether deep learning is a hope for bridging the semantic gap in CBIR and how much empirical improvements can be achieved for learning feature representations and similarity measures, Wan. et al. [79] investigated a framework of deep learning with application to CBIR tasks with an extensive set of empirical studies by examining a state-of-the-art deep convolutional neural network for CBIR tasks under varied settings. Sun et al. [80] proposed a CNN-based image retrieval approach using Siamese network to learn a CNN model for image feature extraction. They used a contrastive loss function to enhance the discriminability of output features. Zagoruyko et al. [81] proposed a general similarity function for patches based on CNN model for learning directly from raw image pixels.

C. SPATIAL TEXTUAL SEARCH

Spatial textual search has been well studied for several years since this technique is significant to local-based services and advanced search engines. It aims to efficiently retrieve a set of spatial textual objects that have a high textual similarity to query keywords and are close enough to query location. Existing literatures show that there are several types of spatial textual search, such as top-k search, k-nearest-neighbor query, range search query, etc.

A wide range of works have been conducted focus on spatial textual search and many solutions have been proposed to improve the system performance. R-Tree is one of the most significant spatial indexing techniques proposed by Guttman [82], which uses minimum bounding area (MBR) to partition the geographical space. Cao et al. [83] studied the problem of collective spatial keyword querying. They proved that the two variants of this problem are NP-complete. For location-aware top-k text retrieval, Cong et al. [50] presented a new indexing framework that integrates the inverted file for text retrieval and the R-tree for spatial proximity querying. Li et al. [84] proposed a novel indexing technique named BR-tree by integrating a spatial component and a textual component to solve the problem of keyword-based kNN search in spatial databases. Based on Quadtree, Zhang et al. [46] proposed a scalable integrated inverted index named I^3. Furthermore, they proposed a novel storage mechanism to improve the efficiency of retrieval and preserve summary information for pruning. To boost the performance of top-k spatial keyword queries, João B. Rocha-Junior et al. [85] designed a novel index named spatial inverted index (S2I) that maps each distinct term to a set of objects containing the term. Li et al. [49] introduced an index structure named IR-Tree which indexes both the textual and spatial contents of documents to support document retrieval and then designed a top-k document search algorithm. Zhang et al. [86] proposed an effective approach to solve the top-k distance-sensitive spatial keyword query by modeling it as the well-known top-k aggregation problem. Zhang et al. [87] introduced a new spatial keyword query problem called m-closest keywords (mCK) query which aims to search out the spatially closest tuples that match m user-specified keywords. To speed up the search, they designed a novel index called the bR*-tree that is extended from R*-tree [88]. Moreover, They exploited a priori-based search strategy to effectively reduce the search space. For collective spatial keyword query problem, Long et al. [89] proposed a distance owner-driven method including an exact algorithm that defeats the best-known existing algorithm and an approximate algorithm which improves the constant approximation factor from 2 to 1.375. For top-k spatial keyword search problem, Zhang et al. [47] presented

an advanced index structure named inverted linear quadtree (IL-Quadtree) to improve efficiency dramatically.

Obviously, these solutions aforementioned just only consider the situation that the geo-location objects containing only one modality data, i.e., text or keywords. In other words, These methods cannot be directly applied to spatial cross-modal retrieval in the geo-multimedia database. This necessitates the development of novel and efficient cross-modal search methods for geo-multimedia data. To the best of our knowledge, this it the first work to imvestigate the problem of geo-multimedia cross-modal retrieval considering both different features of multimodality data and the geographical information.

III. PRELIMINARY

In this section, we firstly formulate the definition of the geo-multimedia object and some relevant notions, then the definition of kNN geo-multimedia cross-modal query is proposed for the first time. Furthermore, we introduce the concept of cross-modal semantic representation mapping. Table 1 summarizes the mathematical notations used throughout this paper to facilitate the discussion of our work.

A. PROBLEM DEFINITION

*Definition 1 (**Geo-Multimedia Object**):* A geo-multimedia objects database is defined as $\mathcal{O} = \{o_1, o_2, ..., o_{|\mathcal{O}|}\}$, wherein $|\mathcal{O}|$ represents the number of objects in $\mathcal{O}$. Each geo-multimedia object $o \in \mathcal{O}$ is associated with a geographical information descriptor $o.\lambda$ and a modality content descriptor $o.M$. A geographical information descriptor includes a 2-dimensional geographical location with longitude X and latitude Y is denoted by $o.\lambda = (X, Y)$. Let $\mathcal{M}$ be the modality set. In this paper we consider two most common modalities, i.e., text and image, thus $\mathcal{M} = \{\mathcal{T}, \mathcal{I}\}$, where $\mathcal{T}$ represents text modality and $\mathcal{I}$ represents image modality. If a geo-multimedia object contains a text, it is denoted as $o.M_{\mathcal{T}}$. Similarly, If an object contains an image, it is denoted as $o.M_{\mathcal{I}}$. $M_{\mathcal{T}}$ and $M_{\mathcal{I}}$ denote the feature vector generated by a text and an image respectively. Let $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$ be the feature spaces of text and image, $\forall o_i \in \mathcal{O}$, if o_i contains a text, then $o_i.M_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}$. If o_i contains an image, then $o_i.M_{\mathcal{I}} \in \mathbb{S}_{\mathcal{I}}$.

Based on the definition of geo-multimedia objects, we define the kNN geo-multimedia cross-modal query. Firstly, we consider the query without geographical information. In other words, we give the definition of cross-modal query and then extend it to the query in the geo-multimedia database.

*Definition 2 (**Coss-Modal Query**):* Given a multimedia objects database $\mathcal{O} = \{o_1, o_2, ..., o_{|\mathcal{O}|}\}$, in which each object contains one of the following two modalities, i.e., text modality $\mathcal{T}$ and image modality $\mathcal{I}$. There are two types of cross-modal query can be defined: (1) $\mathcal{Q}_{\mathcal{T}2\mathcal{I}}$ is defined as a text query which aims to search our the most relevant multimedia object $o \in \mathcal{O}$ contains an image, and $\mathcal{Q}_{\mathcal{T}2\mathcal{I}}.M_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}$,$o.M_{\mathcal{I}} \in \mathbb{S}_{\mathcal{I}}$. (2)$\mathcal{Q}_{\mathcal{I}2\mathcal{T}}$ is defined as a image query which aims to search out the most relevant multimedia object $o \in \mathcal{O}$ contains a text, and $\mathcal{Q}_{\mathcal{I}2\mathcal{T}}.M_{\mathcal{I}} \in \mathbb{S}_{\mathcal{I}}$,$o_i.M_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}$.

Notation	Definition
$\mathcal{O}$	A given database of geo-multimedia objects
$\|\mathcal{O}\|$	The number of objects in $\mathcal{O}$
$o.\lambda$	The geo-location information descriptor of o
$o.\psi$	A visual content descriptor of o
$\mathcal{Q}^k$	A kNN geo-multimedia cross-modal query
$\mathcal{Q}_{\mathcal{T}2\mathcal{I}}$	A text query to search images
$\mathcal{Q}_{\mathcal{I}2\mathcal{T}}$	A image query to search texts
$\mathcal{M}$	A modality set
T	Text modality
I	Image modality
$\mathbb{S}_{\mathcal{T}}$	A text feature space
$\mathbb{S}_{\mathcal{I}}$	A image feature space
$M_{\mathcal{T}}$	a feature vector of a text
$M_{\mathcal{I}}$	a feature vector of an image
X	The longitude of a geo-location
Y	The latitude of a geo-location
k	The number of final results
$\mathcal{F}_{score}(\mathcal{Q}, o)$	The score function
$\mathcal{R}$	The set of results
μ	A parameter to balance distance proximity and semantic similarity
$Dst(\mathcal{Q}, o)$	The spatial distance function.
$\delta(\mathcal{Q}, o)$	The Euclidean distance between $\mathcal{Q}$ and o
$Sim(\mathcal{Q}, o)$	The semantic similarity between $\mathcal{Q}$ and o
$\mathbb{W}_{\mathcal{T}}$	An intermediate representation space of text modality
$\mathbb{W}_{\mathcal{I}}$	An intermediate representation space of image modality
$\mathbb{R}_{\mathcal{T}}$	The semantic representation space of text modality
$\mathbb{R}_{\mathcal{I}}$	The semantic representation space of image modality
Ψ	A mapping from text feature space to image feature space
$\mathfrak{W}$	A cross-modal semantic representation space
$\mathfrak{L}_{\mathcal{T}}$	A non-linear transformation for text modality
$\mathfrak{L}_{\mathcal{I}}$	A non-linear transformation for image modality
$\Theta_{\mathcal{T}}$	A projection from text feature space to intermediate representation space
$\Theta_{\mathcal{I}}$	A projection from image feature space to intermediate representation space
$\mathcal{C}$	The set of semantic concepts
Υ	The set of classes
$\boldsymbol{\iota}$	A visual feature vector of a geo-image I
$\boldsymbol{\tau}$	A text feature vector of a geo-text T
N_i	A node of GMR-Tree
S_i	A signature
$\mathcal{H}_{SIG}(.)$	A hashing function to generate signatures

TABLE 1: The summary of notations

*Definition 3 (k**NN Geo-Multimedia Cross-Modal Query**):* Given a geo-multimedia objects database $\mathcal{O} = \{o_1, o_2, ..., o_{|\mathcal{O}|}\}$, a kNN Geo-Multimedia Cross-Modal Query $\mathcal{Q}^k = (\lambda, M)$ aims to return k nearest geo-multimedia objects whose modalities features are highly relevant to the query. Like Definition 3, we define these two types of query as $\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}$ and $\mathcal{Q}^k_{\mathcal{I}2\mathcal{T}}$, which are named kNN geo-multimedia text to image query (kT2IQ) and kNN geo-multimedia image to text query (kI2TQ) respectively. In more detail, $\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}$ aims to return k nearest geo-multimedia objects which contain images that

are highly relevant to the query text, and $\mathcal{Q}^k_{\mathcal{I}2\mathcal{T}}$ aims to find k nearest objects which contain texts that are highly relevant to the query image. The relevancy between text and image is the semantic correlation between them. Formally, For query $\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}$, the result is k geo-multimedia objects $R_{\mathcal{T}2\mathcal{I}}$ which are ranked by the a score function $\mathcal{F}_{score}(\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}, o)$,i.e.,

$$\begin{aligned}\mathcal{R}_{\mathcal{T}2\mathcal{I}} = \{o|\forall o \in \mathcal{O}, o' \in \mathcal{O} \setminus \mathcal{R}_{\mathcal{T}2\mathcal{I}},\\ \mathcal{F}_{score}(\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}, o) > \mathcal{F}_{score}(\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}, o')\},\\ \mathcal{R}_{\mathcal{T}2\mathcal{I}} \subseteq \mathcal{O}, |\mathcal{R}_{\mathcal{T}2\mathcal{I}}| = k\end{aligned} \tag{1}$$

likewise, for query $\mathcal{Q}^k_{\mathcal{I}2\mathcal{T}}$, the result is k geo-multimedia objects $R_{\mathcal{T}2\mathcal{I}}$ ranked by $\mathcal{F}_{score}(\mathcal{Q}^k_{\mathcal{I}2\mathcal{T}}, o)$, i.e.,

$$\begin{aligned}\mathcal{R}_{\mathcal{I}2\mathcal{T}} = \{o|\forall o \in \mathcal{O}, o' \in \mathcal{O} \setminus \mathcal{R}_{\mathcal{I}2\mathcal{T}},\\ \mathcal{F}_{score}(\mathcal{Q}^k_{\mathcal{I}2\mathcal{T}}, o) > \mathcal{F}_{score}(\mathcal{Q}^k_{\mathcal{I}2\mathcal{T}}, o')\},\\ \mathcal{R}_{\mathcal{I}2\mathcal{T}} \subseteq \mathcal{O}, |\mathcal{R}_{\mathcal{I}2\mathcal{T}}| = k\end{aligned} \tag{2}$$

and the score function is defined as follows:

$$\mathcal{F}_{score}(\mathcal{Q}, o) = \mu Dst(\mathcal{Q}, o) + (1-\mu)Sim(\mathcal{Q}.o) \tag{3}$$

where $\mathcal{Q}$ represents a query, and $\mu \in [0, 1]$ is a parameter which is to balance the importance between distance proximity component and semantic similarity component. If $\mu > 0.5$, it means the distance proximity is more important than the semantic similarity. And if $\mu = 0$, it means this function is just used to measure the semantic similarity between $\mathcal{Q}$ and o.

In this paper, we focus on the kT2IQ query $\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}$: given a query text, the system will measure the geographical distance proximity according the geo-locations of query and objects, and meanwhile measure the relevance between query text and images contained in objects. To facilitate the expression, we abbreviate $\mathcal{Q}^k_{\mathcal{T}2\mathcal{I}}$ as $\mathcal{Q}$. In the following part we introduce how to measure spatial distance proximity and the cross-modal semantic correlation.

*Definition 4 (**Spatial distance proximity measurement**):* Given a geo-multimedia objects database $\mathcal{O} = \{o_1, o_2, ..., o_{|\mathcal{O}|}\}$ and a kT2IQ query $\mathcal{Q}$, $\forall o \in \mathcal{O}$, the spatial distance proximity is measured by the following function:

$$Dst(\mathcal{Q}, o) = 1 - \frac{\delta(\mathcal{Q}, o)}{\delta_{max}(\mathcal{Q}, \mathcal{O})} \tag{4}$$

where $\delta(\mathcal{Q}, o)$ represents Euclidean distance between the query $\mathcal{Q}$ and the object o. $\delta_{max}(\mathcal{Q}, \mathcal{O})$ represents the maximum spatial distance between $\mathcal{Q}$ and any objects in $\mathcal{O}$. They are defined in detail as follows:

$$\delta(\mathcal{Q}, o) = \sqrt{(\mathcal{Q}.\lambda.X - o.\lambda.X)^2 + (\mathcal{Q}.\lambda.Y - o.\lambda.Y)^2} \tag{5}$$

$$\delta_{max}(\mathcal{Q}, \mathcal{O}) = max(\{\delta(\mathcal{Q}, o)|\forall o \in \mathcal{O}\}) \tag{6}$$

where the function $max(\mathcal{X})$ is to return the maximum value of element in the set $\mathcal{X}$. It is easily to know that for spatial distance proximity measurement, the objects with the **small score values** are preferred (i,e., ranked higher).

*Definition 5 (**Cross-modal semantic similarity measurement**):* Given a geo-multimedia objects database $\mathcal{O} = \{o_1, o_2, ..., o_{|\mathcal{O}|}\}$ and a kT2IQ query $\mathcal{Q}$, $\forall o \in \mathcal{O}$, the cross-modal semantic similarity is measured by cosine similarity measurement, as shown in the following equation:

$$\begin{aligned}&Sim(\mathcal{Q}, o)\\ &= \frac{\sum_{i \in \mathcal{Q}.M_{\mathcal{T}}} \mathcal{Q}.M^{(i)}_{\mathcal{T}} * o.M^{(i)}_{\mathcal{I}}}{\sqrt{\sum_{i \in \mathcal{Q}.M_{\mathcal{T}}} (\mathcal{Q}.M^{(i)}_{\mathcal{T}})^2} * \sqrt{\sum_{i \in o.M_{\mathcal{I}}} (o.M^{(i)}_{\mathcal{I}})^2}}\end{aligned} \tag{7}$$

where $\mathcal{Q}.M^{(i)}_{\mathcal{T}}$ and $o.M^{(i)}_{\mathcal{I}}$ represent ith feature element in representation vector $\mathcal{Q}.M_{\mathcal{T}}$ and $o.M_{\mathcal{I}}$ respectively.

B. CROSS-MODAL SEMANTIC REPRESENTATION SPACE

It is common knowledge that semantic gap is a ticklish problem for cross-modal retrieval. In other words, we cannot directly measure similarity between query and object which belongs to different modalities by equation (7). Because $\mathcal{Q}.M_{\mathcal{I}}$ and $o.M_{\mathcal{I}}$ cannot be mapped into a common space. Therefore, this task cannot be reduced to a classical information retrieval task in which there is a mapping between query representation space and object representation space. It can be described in formal as follows: for a query $\mathcal{Q}$ with a text and a geo-multimedia object o with an image, the features spaces of them are denoted as $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$ respectively, and $\mathcal{Q}.M_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}, o.M_{\mathcal{I}} \in \mathbb{S}_{\mathcal{I}}$, the mapping between $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$ is represented as

$$\Psi : \mathbb{S}_{\mathcal{T}} \longrightarrow \mathbb{S}_{\mathcal{I}}$$

and the inverse mapping is represented as

$$\Psi^{-1} : \mathbb{S}_{\mathcal{I}} \longrightarrow \mathbb{S}_{\mathcal{T}}$$

Thus, the cross-modal text to image query can be denoted as $\mathcal{Q}_{\mathcal{T}2\mathcal{I}} \iff \Psi(\mathcal{Q}.M_{\mathcal{T}})$. As discussed above, it is hard to find this mapping between feature spaces of different modalities.

To this end, we assume that there exist two mappings which map text and image feature spaces into two intermediate representation $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$ respectively, that is:

$$\Omega_{\mathcal{T}} : \mathbb{S}_{\mathcal{T}} \longrightarrow \mathbb{W}_{\mathcal{T}}$$

$$\Omega_{\mathcal{I}} : \mathbb{S}_{\mathcal{I}} \longrightarrow \mathbb{W}_{\mathcal{I}}$$

and the inverse mappings of them are denoted respectively as

$$\Omega^{-1}_{\mathcal{T}} : \mathbb{W}_{\mathcal{T}} \longrightarrow \mathbb{S}_{\mathcal{T}}$$

$$\Omega^{-1}_{\mathcal{I}} : \mathbb{W}_{\mathcal{I}} \longrightarrow \mathbb{S}_{\mathcal{I}}$$

and existing a mapping Φ:

$$\Phi : \mathbb{W}_{\mathcal{T}} \longrightarrow \mathbb{W}_{\mathcal{I}}$$

that means there is a semantic correlation between these two isomorphic spaces $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$.

According to this assumption, we redescribe the cross-modal text to image query in the following forms: Given a geo-multimedia database $\mathcal{O}$, a kT2IQ query $\mathcal{Q}$ is to search

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

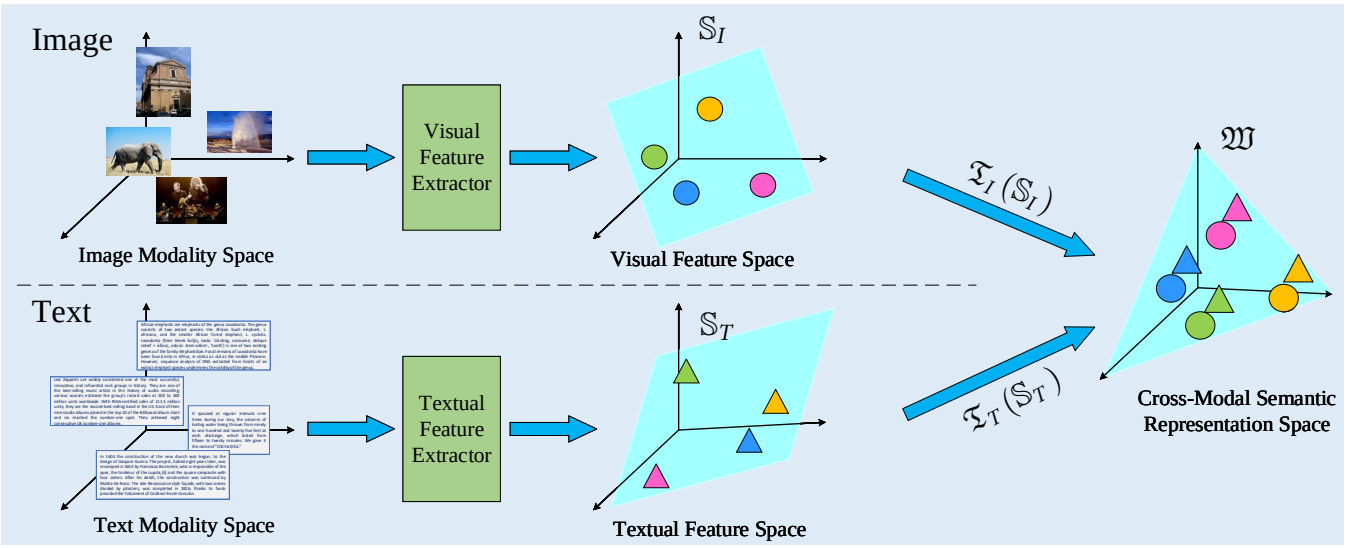


FIGURE 3: The construction of cross-modal semantics representation space. Herein we only consider geo-image and geo-text modalities. Two feature extractors map the geo-multimedia objects from original modality space to feature space, and the feature vectors of image and text are transformed into semantic representations in cross-modal semantic representations space via two non-linear mappings.

out the most relevant object contains image that is represented as $\Omega_{\mathcal{I}}^{-1}(\Phi(\Omega_{\mathcal{T}}(\mathcal{Q}.M_{\mathcal{T}})))$ in $\mathbb{S}_{\mathcal{I}}$. In other words, This idea is to use two intermediate representation spaces $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$ to implement the mapping from $\mathbb{S}_{\mathcal{T}}$ to $\mathbb{S}_{\mathcal{I}}$.

According to the above discussion, the most difficult problem for implementing efficient cross-modal retrieval is to learn the intermediate representation spaces $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$. To overcome this challenge, we introduce a notion named **CrOss-modal Semantics Representation Space** (**CoSReS**), shown as follows.

*Definition 6 (**Cross-Modal Semantic Representation Space (CoSReS)**):* Given a geo-multimedia database $\mathcal{O}$ and modality set $\mathcal{M} = \{\mathcal{T}, \mathcal{I}\}$. Let $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$ be the feature spaces of text and image respectively, $\mathbb{R}_{\mathcal{T}}$ and $\mathbb{R}_{\mathcal{I}}$ be the semantic space of text and image respectively. A CoSReS $\mathfrak{W}$ is a isomorphic representation space for modalities $\mathcal{T}$ and $\mathcal{I}$ in a high-level semantic abstraction, if existing two non-linear transformations $\mathfrak{T}_{\mathcal{T}}$ and $\mathfrak{T}_{\mathcal{I}}$, $\mathbb{R}_{\mathcal{T}} = \mathfrak{T}_{\mathcal{T}}(\mathbb{S}_{\mathcal{T}})$ and $\mathbb{R}_{\mathcal{I}} = \mathfrak{T}_{\mathcal{I}}(\mathbb{S}_{\mathcal{I}})$, then $\mathfrak{W} = \mathbb{R}_{\mathcal{T}} = \mathbb{R}_{\mathcal{I}}$.

Fig. 3 demonstrates the concept of CoSReS. For two different modalities, CoSReS have a set of common semantic concepts. After extracting features for texts and images respectively, the feature vectors of texts and images can be transformed into semantic representation vectors in CoSReS. Therefore, we can easily measure the semantic similarity in this common representation space.

IV. THE FRAMEWORK

In this section, we propose a novel framework for geo-multimedia cross-modal retrieval, which includes multi-modal feature extraction, cross-modal semantic space mapping, geo-multimedia spatial index and cross-modal semantic similarity measurement. As mentioned above, this framework is desinged for kNN geo-text to geo-image query kT2IQ, but this approach can also be extended for other modalities, e.g. audio and video by changing the feature representation component. In this section, a overview of this framework is given and the details of each component are presented in the next two sections.

Feature Extraction. Specifically, two datasets, as shown in Fig. 4, i.e., geo-image set and geo-text set are used to train the feature extraction models called VisNet and TxtNet for image and text respectively, which generate feature representations. In other words, VisNet and TxtNet play the roles of feature mappings that maps geo-image objects and geo-text objects into visual feature space and text feature space, namely $VisNet(\{I_1, I_2, ...I_m\}; \boldsymbol{\theta}) = \{\boldsymbol{\iota}_1, \boldsymbol{\iota}_2, ..., \boldsymbol{\iota}_m\}$, $TxtNet(\{T_1, T_2, ...T_m\}; \boldsymbol{\psi}) = \{\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, ..., \boldsymbol{\tau}_m\}$, where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the model parameters of VisNet and TxtNet. Apparently, there are several ways to implement VisNet and TxtNet, such as SIFT, BoW, LDA in a traditional manner, or CNN and LSTM in a deep learning based manner. In this work we employ AlexNet and LDA model to implement VisNet and TxtNet, which are explained minutely in Section V. Other techniques will be exploited in our future works.

Semantic Representation. As discussed above, the main obstacle of the cross-modal retrieval problem is the semantic gap between different modalities. How to bridge the semantic gap is one of the main challenges of cross-modal retrieval task. To this end, we propose to construct a cross-modal semantic representation space in which different modalities objects can be represented by common highe-level semantic concepts. In other words, the semantic similarity between these cross-modal objects can be easily measured precisely in a traditional way (e.g., cosine similarity). We propose a novel method named **Cro**ss-modal **S**emantic **Mat**ching

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

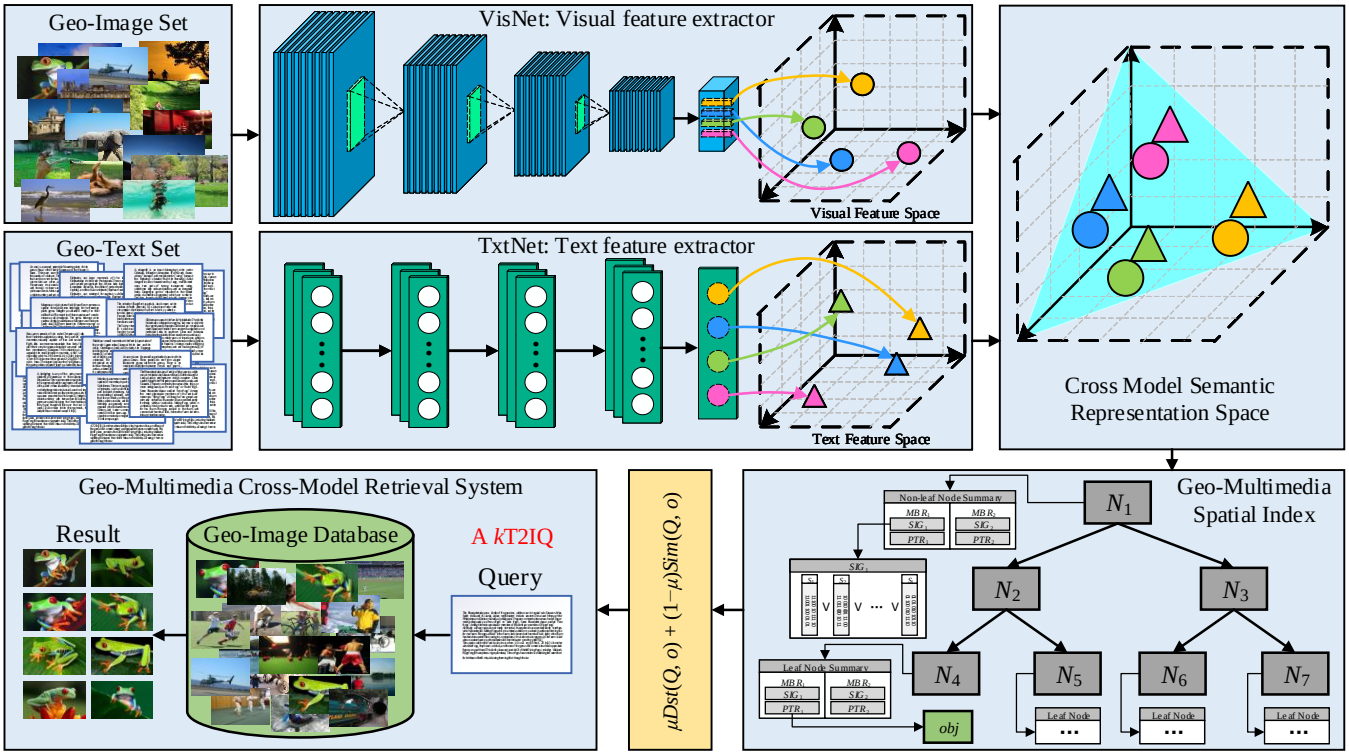


FIGURE 4: The proposed framework for geo-multimedia cross-modal retrieval. It is designed for kNN geo-multimedia text to image query kT2IQ. Two feature extractors, namely VisNet and TxtNet, which are learning based methods to extract visual features and text features from geo-images and geo-texts, respectively. In other words, they map geo-images and geo-texts into visual feature space and text feature space. To overcome the challenge of semantic gap between image modality and text modality, we propose to construct a corss-modal semantic representation space in which we can measure the semantic similarity between the semantic representations of geo-images and geo-texts. Based on the cross-modal semantic representations, a novel hybrid index that is a combination of R-Tree and signature files is carefully designed and an efficient kNN geo-multimedia cross-modal search algorithm is developed to speed up the retrieval. Aaccording to the score function $\mathcal{F}_{score}(\mathcal{Q}, o) = \mu Dst(\mathcal{Q}, o) + (1-\mu)Sim(\mathcal{Q}.o)$, the system can measure the similarity between query Q and an geo-multimedia object o in both aspects of geo-location and semantic concept precisely.

(**CoSMat**) consists of two novel techniques, namely CorrProj and LogsTran to implement non-linear mappings from feature space to semantic space. This method is described in Section V in detail.

Spatial Indexing. To boost the efficiency of the large-scale geo-multimedia retrieval, we propose to develop a hybrid spatial index structure and integrate it into this framework. Inspired by traditional spatial textual search techniques, i.e., R-Tree and signature method, an exquisitely designed index structure named GMR-Tree is proposed, in which the cross-modal semantic representations in CoSReS are used to generate signature files in binary and stored in the tree nodes. Similar to R-Tree, the geo-location informantion such as longitude and latitude are used to partition the geographical space in the form of minimum bounding area (MBR). This part is detailed discussed in Section VI.

Similarity Measurement and Search. Based on GMR-Tree, we design a kNN geo-multimedia cross-modal search algorithm, called kGMCMS. The score function $\mathcal{F}_{score}(\mathcal{Q}, o) = \mu Dst(\mathcal{Q}, o) + (1-\mu)Sim(\mathcal{Q}.o)$ defined in Section III is used to measure the similarity between the query Q and the geo-multimedia object o in both aspects of geographical proximity and semantic correlation. The implementation of this algorithm is introduced in Section VI.

V. CROSS-MODAL SEMANTIC REPRESENTATION SPACE CONSTRUCTION WITH DEEP LEARNING

In this section, we reduce the task of bridging the semantic gaps between different modalities into the problem of intermediate representation space construction, which can be represented by cross-modal semantic representation space (CoSReS). In this section, we present a deep learning based solution to construct the CoSReS based on the concept presented in subsection III-B. First we discuss how to learn a common semantic representation space for text and image data. Then an effective approach named DeCoSReS is introduced, which utilizes convolution neural networks (CNN) and Latent Dirichlet Allocation [91] (LDA) to learn the representation speace.

A. CROSS-MODAL SEMANTIC MATCHING

We use the method called cross-modal semantic matching (CoSMat) to construct CoSReS so that it provides a common

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

semantic representation space for different modalities. This algorithm consists of two components, i.e., (1)CCA based **Corr**elation **Proj**ection (**CorrProj**) and (2)**logis**tic regression based **Tran**sformation (**LogsTran**). The former aims to learn subspaces from feature spaces of different modalities, and the latter is to learn semantic mappings in these subspaces. We introduce these two important techniques respectively in the following part.

CorrProj. Canonical correlation analysis [90] (CCA) is a popular dimensionality reduction method. We use it to learn γ-dimensional subspaces $\mathbb{W}_{\mathcal{T}}^{\gamma} \in \mathbb{S}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}^{\gamma} \in \mathbb{S}_{\mathcal{I}}$ to find the correlations between these two subspaces. CCA method learns directions in text and image feature spaces, i.e., $\Gamma_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}$ and $\Gamma_{\mathcal{I}} \in \mathbb{S}_{\mathcal{I}}$ along the directions of the data maximally correlated. That is, for feature vectors $M_{\mathcal{T}}$ and $M_{\mathcal{I}}$, measuring the maximun correlation:

$$\mathfrak{u} = \Gamma_{\mathcal{T}}^T M_{\mathcal{T}},$$

$$\mathfrak{v} = \Gamma_{\mathcal{I}}^T M_{\mathcal{I}},$$

$$maxCorr(\mathfrak{u}, \mathfrak{v}) = \frac{\Gamma_{\mathcal{T}}^T \Sigma_{\mathcal{T}\mathcal{I}} \Gamma_{\mathcal{I}}}{\sqrt{\Gamma_{\mathcal{T}}^T \Sigma_{\mathcal{T}\mathcal{T}} \Gamma_{\mathcal{T}}} \sqrt{\Gamma_{\mathcal{I}}^T \Sigma_{\mathcal{I}\mathcal{I}} \Gamma_{\mathcal{I}}}} \quad (8)$$

wherein $\Sigma_{\mathcal{T}\mathcal{T}}$ and $\Sigma_{\mathcal{I}\mathcal{I}}$ are the empirical covariance matrices of space $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$, i.e., $\Sigma_{\mathcal{T}\mathcal{T}} = Cov(\mathbb{S}_{\mathcal{T}})$ and $\Sigma_{\mathcal{I}\mathcal{I}} = Cov(\mathbb{S}_{\mathcal{I}})$, $\Sigma_{\mathcal{T}\mathcal{I}}$ is the empirical cross-covariance matrix of them, i.e., $\Sigma_{\mathcal{T}\mathcal{I}} = Cov(\mathbb{S}_{\mathcal{T}}, \mathbb{S}_{\mathcal{I}})$, and $\Sigma_{\mathcal{T}\mathcal{I}} = \Sigma_{\mathcal{I}\mathcal{T}}^T$.

The first γ canonical components $\{\Gamma_{\mathcal{T}_1}\}^{\gamma}$ and $\{\Gamma_{\mathcal{I}_1}\}^{\gamma}$ represent a basis for projection $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$ on subspace $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$. For each text $M_{\mathcal{T}}$ in space $\mathbb{S}_{\mathcal{I}}$, it can be mapped into the projection $\Theta_{\mathcal{T}}(M_{\mathcal{T}})$ onto $\{\Gamma_{\mathcal{T}_1}\}^{\gamma}$. Likewise, for each image $M_{\mathcal{I}}$ in space $\mathbb{S}_{\mathcal{I}}$, it can be mapped into the projection $\Theta_{\mathcal{I}}(M_{\mathcal{I}})$ onto $\{\Gamma_{\mathcal{I}_1}\}^{\gamma}$. Therefore, the method CorrProj can learn two projections $\Theta_{\mathcal{T}}(M_{\mathcal{T}})$ and $\Theta_{\mathcal{I}}(M_{\mathcal{I}})$ from $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$, which can be used to define two γ-dimension subspaces for text and image, i.e.,

$$\Theta_{\mathcal{T}} : \mathbb{S}_{\mathcal{T}} \longrightarrow \mathbb{W}_{\mathcal{T}}$$

and,

$$\Theta_{\mathcal{I}} : \mathbb{S}_{\mathcal{I}} \longrightarrow \mathbb{W}_{\mathcal{I}}$$

After that, this approach used another component named LogsTran to learn two semantic mappings from these two subspace, which is described as follows.

LogsTran. The method aforementioned is to map feature spaces of text and image to maximally correlated subspaces $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$. Then we use another method called LogsTran to find the correspondence between $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$ by represented objects at a higher-level of semantic abstraction. It can map text and image space into a common semantic representation space with a set of semantic concepts $\mathcal{C} = \{c_1, c_2, ..., c_n\}$, such as "airplane","cat" or "house". We utilize logistic regression to learn two transformation $\mathfrak{L}_{\mathcal{T}}$ and $\mathfrak{L}_{\mathcal{I}}$. $\mathfrak{L}_{\mathcal{T}}$ transforms a text contained by a geo-multimedia object $o.M_{\mathcal{T}} \in \mathbb{S}_{\mathcal{T}}$ into a vector of posterior probabilities $P_{\mathcal{T}}^{\Upsilon}(v_i|\mathcal{T})$, in which $\Upsilon = \{v_1, v_2, ..., v_k\}$ is a set of classes. Likewise, $\mathfrak{L}_{\mathcal{I}}$ transforms an image contained by a geo-multimedia object $o.M_{\mathcal{I}} \in \mathbb{S}_{\mathcal{I}}$ into a vector of posterior probabilities $P_{\mathcal{I}}^{\Upsilon}(v_i|\mathcal{I})$. The spaces $\mathbb{R}_{\mathcal{T}}$ and $\mathbb{R}_{\mathcal{I}}$ of these posterior probabilities vectors are referred to the semantic representation space of text and image respectively. Formally, they can be presented as follows:

$$\mathfrak{L}_{\mathcal{T}} : \mathbb{S}_{\mathcal{T}} \longrightarrow \mathbb{R}_{\mathcal{T}}$$

$$\mathfrak{L}_{\mathcal{I}} : \mathbb{S}_{\mathcal{I}} \longrightarrow \mathbb{R}_{\mathcal{I}}$$

Multi-class logistic regression is utilized, which produces a linear classifier. It calculates the posterior probability of class c_i by the following logistic function:

$$P_M^{\Upsilon}(c_i|M_x; \varpi) = \frac{1}{\sum_{c_i} exp(\varpi_{c_i}^T M_x)} exp(\varpi_{c_i}^T M_x) \quad (9)$$

where M represents the modalities information. For example, for text, $M = \mathcal{T}$ and for image , $M = \mathcal{I}$. M_x is the features vector in the input space. $\varpi = (\varpi_1, \varpi_2, ..., \varpi_k)$ is a vector of parameters for class c_i.

According to the logistic regression, in semantic representation spaces $\mathbb{R}_{\mathcal{T}}$ and $\mathbb{R}_{\mathcal{I}}$, the features are semantic concept probabilities, for instance, the probability of a text belongs to "cat" class or the probability of an image belongs to "airplane" class. Furthermore, texts and images are represented as posterior probabilities vectors in regard to same classes. In addition, the semantic representation spaces $\mathbb{R}_{\mathcal{T}}$ and $\mathbb{R}_{\mathcal{I}}$ are isomorphic, and they can be regarded as the same, i.e., $\mathbb{R}_{\mathcal{T}} = \mathbb{R}_{\mathcal{I}}$. Therefore, the cross-modal semantic representation space $\mathfrak{W} = \mathbb{R}_{\mathcal{T}} = \mathbb{R}_{\mathcal{I}}$.

The CosMat method is a combination of CorrProj and LogsTran. In the first step, CorrProj is applied to learn two maximally correlated subspaces $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$ based on feature spaces $\mathbb{S}_{\mathcal{T}}$ and $\mathbb{S}_{\mathcal{I}}$. Then LogsTran method is used to generate two transformations $\mathfrak{L}_{\mathcal{T}}$ and $\mathfrak{L}_{\mathcal{I}}$ to create the isomorphic semantic representation spaces $\mathbb{R}_{\mathcal{T}}$ and $\mathbb{R}_{\mathcal{I}}$. Thus, we can measure the semantic similarity of text and image in the CoSReS $\mathfrak{W}$, i.e., $Sim(\xi_{\mathcal{T}}, \xi_{\mathcal{I}})$, where $\xi_{\mathcal{T}} = \mathfrak{L}_{\mathcal{T}}(\Theta_{\mathcal{T}}(\mathbb{S}_{\mathcal{T}}))$, $\xi_{\mathcal{I}} = \mathfrak{L}_{\mathcal{I}}(\Theta_{\mathcal{I}}(\mathbb{S}_{\mathcal{I}}))$. It is an significant step of implementing kT2IQ.

B. CROSS-MODAL SEMANTIC REPRESENTATION SPACE LEARNING

Deep learning techniques such as CNN, RNN, etc. are widely applied in the area of multimedia retrieval. To implement cross-modal semantic representation space construction and cross-modal retrieval, we employ AlexNet and LDA model to implement VisNet and TxtNet respectively. Fig. 5 is the deep learning based framework of cross-modal semantic representation space construction.

VisNet. For visual features extraction, we use the pre-trained CNN model, AlexNet, proposed by [29] in this framework. It contains five convolutional layers and two fully-connected layers, trained by 1 million images. Specifically, each image is resized to 256×256 at first and then put into this model. The first convolutional layer filters the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

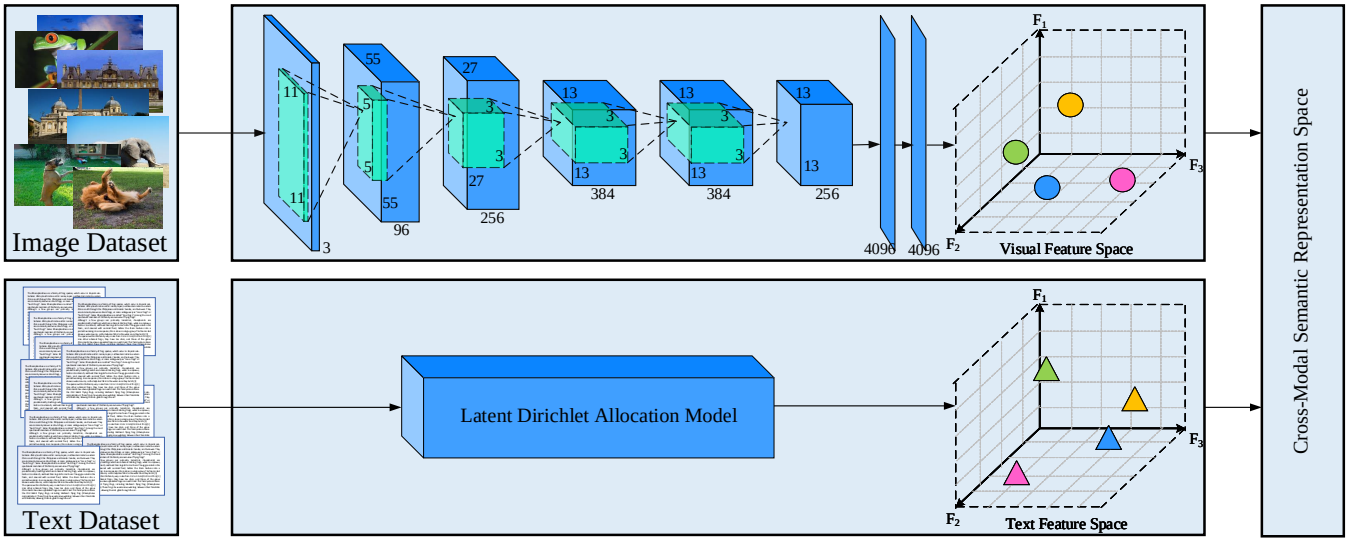


FIGURE 5: The deep learning based framework of cross-modal semantic representation space construction. The VisNet is implemented by AlexNet and the TxtNet is implemented by a LDA Model.

$224 \times 224 \times 3$ input image, which has 96 kernels of size $11 \times 11 \times 3$. The second convolutional layer has 256 kernels of size $5 \times 5 \times 96$. The third convolutional layer has 384 kernels of size $3 \times 3 \times 256$. The fourth convolutional layer has 384 kernels of size $3 \times 3 \times 192$. The fifth convolutional layer has 256 kernels with size of $3 \times 3 \times 192$. The fully-connected layers have 4096 neurons each, which denote 4096 dimensional features after ReLU. In order to improve the performance of visual information recognition, we fine-tune the network parameters by retraining this model on our experimental dataset, namely Flickr.

TxtNet. For textual feature extraction, we utilize Latent Dirichlet Allocation (LDA) model to generate the representation of the input text. LDA is a generative model for a text corpus in which the semantic content of a text is summarized as a mixture of serval topics. Specifically, a text is modeled by a multinomial distribution over κ topics and each word in a text is generated by first sampling a topic from the text-speccific topic distribution [91].

As the first study of geo-multimedia cross-modal retrieval, we use the simple but effective method (AlexNet and LDA) for CoSReS learning. Nevertheless, this combination is by no means the only choice. Other powerful deep learning model e.g. VGGNet [92], GoogLeNet [93] and ResNet [94] for image, and RNN [95], BiLSTM [96], [97] for text can also play the role of VisNet and TxtNet. We will investigate these models in our future work.

After generating multi-modal feature representations via VisNet and TxtNet, CorrProj and LogsTran are combined to generate cross-modal semantic representation space $\mathfrak{W}$. Specifically, for image and text, the correlation subspaces $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$ are built by CorrProj from the textual and visual feature vectors. Then, two semantic mappings are learned from $\mathbb{W}_{\mathcal{T}}$ and $\mathbb{W}_{\mathcal{I}}$ by LogsTran. That means $\mathfrak{L}_{\mathcal{T}}$ and $\mathfrak{L}_{\mathcal{I}}$ map the text and image into a common metric space. Therefore, based on these two semantic mapping, the similarity of text and image can be measured.

VI. HYBIRD INDEXING FOR GEO-MULTIMEDIA CROSS-MODAL RETRIEVAL

In this section, we present a novel hybrid spatial indexing technique for efficient geo-multimedia cross-modal retrieval. We call this index **G**eo-**M**ultimedia **R-Tree** (**GMR-Tree**). Firstly we introduce the basic structure of GMR-Tree and related concepts. Then we propose our search algorithm that can boost the performance of geo-multimedia cross-modal query.

A. HYBRID INDEXING STRUCTURE

The proposed hybrid index is called GMR-Tree. It is a combination of an R-Tree [82] and signature files. Different from R-Tree, the nodes of GMR-Tree not only contain geo-location information, but carry modality semantic representation information as well. The geo-location information is represented in the form of minimum bounding area (MBR) and semantic representation information is in the form of a signature. In the following part, we introduce this novel indexing technique in detail.

Fig. 6 illustrates the structure of a GMR-Tree. Generally, a GMR-Tree is a height-balanced tree structure. Each non-leaf node denoted as a triple $\langle MBR, SIG, PTR_N \rangle$ contains three components. *MBR* is defined as in the R-Tree, which represents the geo-location in the form of minimum bounding area (MBR). *SIG* is a signature file generated from the geo-multimedia objects in this *MBR*. For the ith object o_i in *MBR*, its signature is denoted as $S_i = \mathcal{H}_{SIG}(o_i.M_I)$, wherein $\mathcal{H}_{SIG}(.)$ is a hashing function which is used to generate a signature from the semantic representation vector. For a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

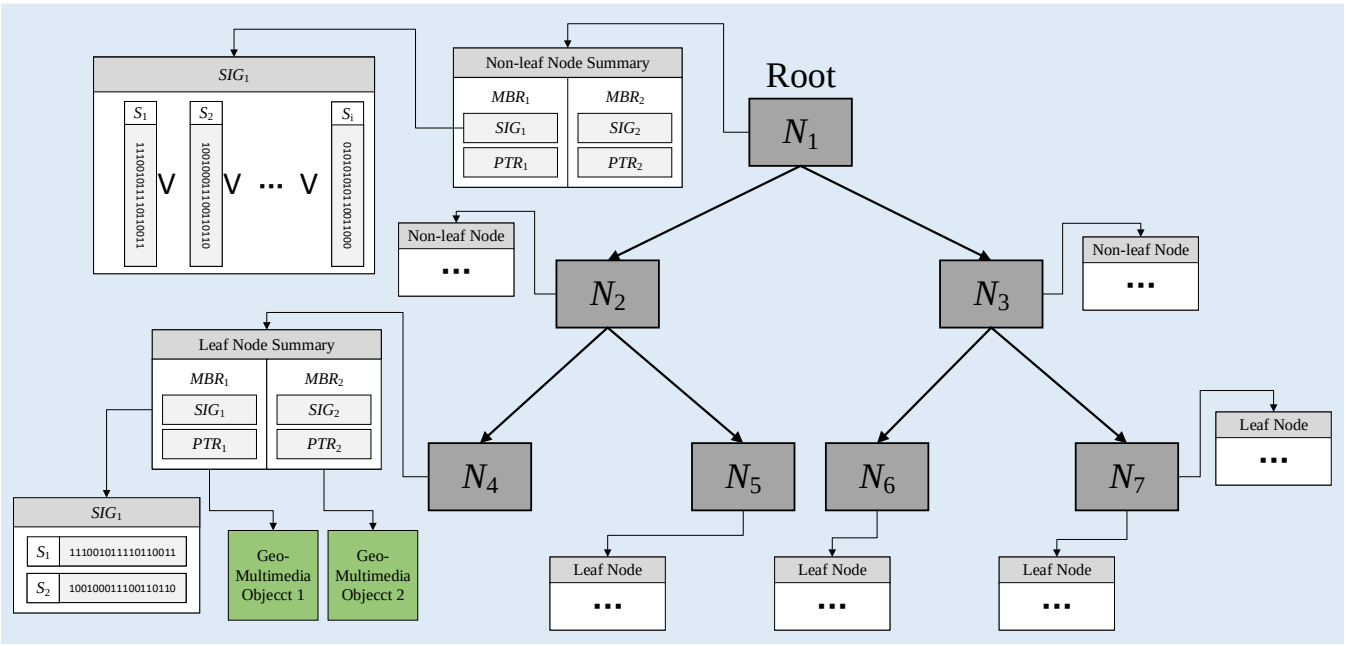


FIGURE 6: A GMR-Tree. It is a combination of R-Tree and signature files. The semantic representations of geo-multimedia objects are stored in the tree nodes and the geographical space is partitioned by MBR.

MBR_1, the signature $SIG_1 = S_1 \bigvee S_2 \bigvee ... \bigvee S_i$, wherein the operator $\bigvee$ represents binary OR-ing operation. In other words, the signature of a node is equivalent to a signature that superimposes the signatures of the children nodes. In addition, the length of the signatures in each level is the same. The third component of node is a pointer PTR_N, which refers to a subnode. Similarly, the leaf note in GMR-Tree is the form of $\langle MBR, SIG, PTR_o \rangle$ but the pointer PTR_o refers to point geo-multimedia objects.

There is a very useful property of GMR-Tree, which can provide well support for the spatial search. We describe it as follows.

Property 1: Given a query $\mathcal{Q}$ and a node N_i, the signatures of $\mathcal{Q}$ and N_i are $SIG_\mathcal{Q}$ and SIG_i respectively. If $SIG_\mathcal{Q} = SIG\mathcal{Q} \bigvee SIG$, that means the query $\mathcal{Q}$ contains some same semantic concepts as the objects in N_i. In other words, the query may be similar to some objects in N_i on semantic level. Otherwise, $\mathcal{Q}$ may be dissimilar to the objects in the node.

B. KNN GEO-MULTIMEDIA CROSS-MODAL SEARCH ALGORITHM

Based on GMR-Tree and its property, we design an efficient spatial search algorithm to support kNN geo-multimedia cross-modal retrieval. The pseudo-code of kGMCMS algorithm is demonstrated in Algorithm 1. Algorithm 2 is the GMR-Tree based nearest neighbor search algorithm that is used in kGMCMS.

For Algorithm 1, in the first step, a priority queue $\mathcal{L}$ is initialized as a empty set and an integer α which is used for counting during the search. $\mathcal{R}$ is the set of results. First the algorithm puts the root node of GMR-Tree $\mathcal{G}$ into $\mathcal{L}$, and then generates the signature for query $\mathcal{Q}$. In this process, each element of semantic representation vector $\mathcal{Q}.M_\mathcal{T}$ is reassigned by a hashing function $\mathcal{H}_{SIG}(.)$ that converts the element of $\mathcal{Q}.M_\mathcal{T}$ into a hash code. After that, the search process is implemented by a While loop. During the process, the nearest neighbor o of query $\mathcal{Q}$ is found out and then the score of o is calculated by score function $\mathcal{F}_{score}(\mathcal{Q}, o)$ which is introduced in section III. Here we set $\mu = 0.5$. That means the geographical distance proximity is same important as semantic correlation.

For Algorithm 2, we initialize a variable $\mathcal{E}$ to store a tree node. $\mathcal{L}$ will be checked circularly whether it is empty or not. If $\mathcal{L}$ is not empty, the algorithm gets a node stored in $\mathcal{L}$ by a $Dequeue(.)$ operation and put it into $\mathcal{E}$. If this node is a non-leaf node, and exist an object whose *SIG* matches the query, then measures the distance between $\mathcal{Q}$ and *MBR* of $\mathcal{E}$. It will be put into $\mathcal{L}$ again. If $\mathcal{E}$ is a leaf node, all objects in it will be checked and put the object which matches the query in to $\mathcal{L}$.

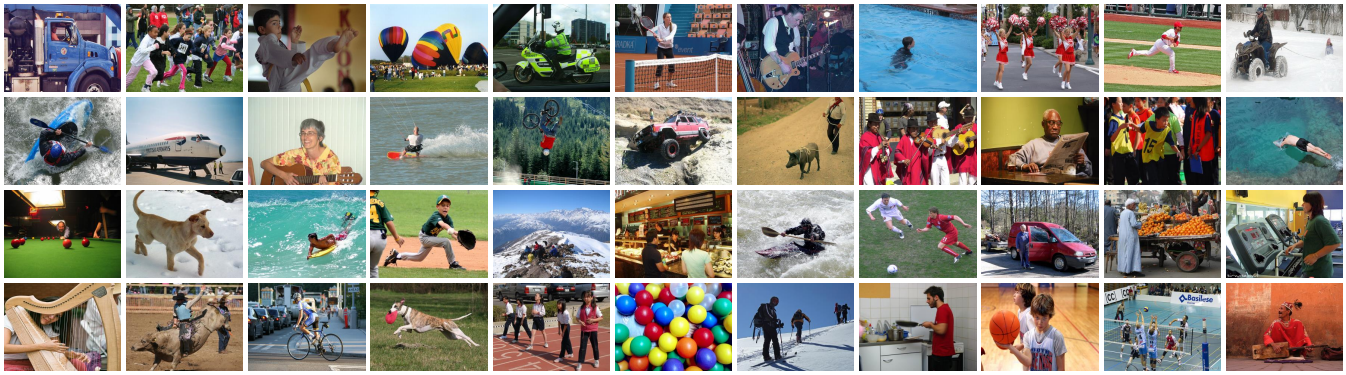
VII. EXPERIMENTAL EVALUATION

In this section, we conduct a comprehensive experiments on a real and a synthetic dataset to evaluate the performance of the proposed method, i.e., **DeCoSReS+GMR-Tree**. Firstly we introduce the datasets and workload in subsection VII-A, and then discuss the evaluations in subsection VII-B.

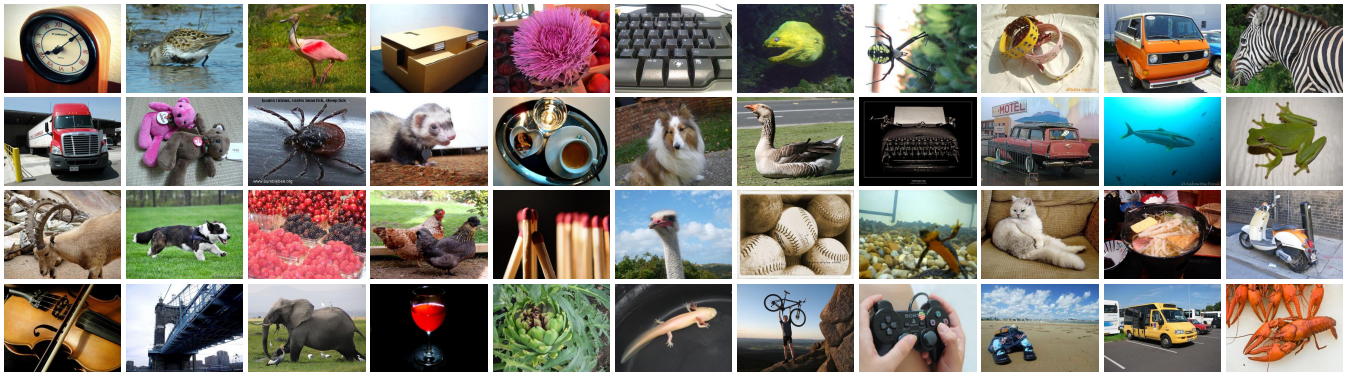
A. DATESET AND WORKLOAD

Dataset. Our experiments aim to evaluate the performance of the proposed approach on a real geo-multimedia dataset and a synthetic dataset:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access



Some samples of Flickr dataset



Some samples of ImageNet dataset

FIGURE 7: The samples of the real and synthetic dataset. The real dataset is a collection of geo-images crawled from Flickr (http://www.flickr.com/), which have geo-location information. The synthetic dataset contains images from ImageNet (http://image-net.org/index), and these images are tagged by the geo-locations from Rtree-Portal (http://www.rtreeportal.org).

Algorithm 1 k**NN Geo-Multimedia Cross-Modal Search (**k**GMCMS)**

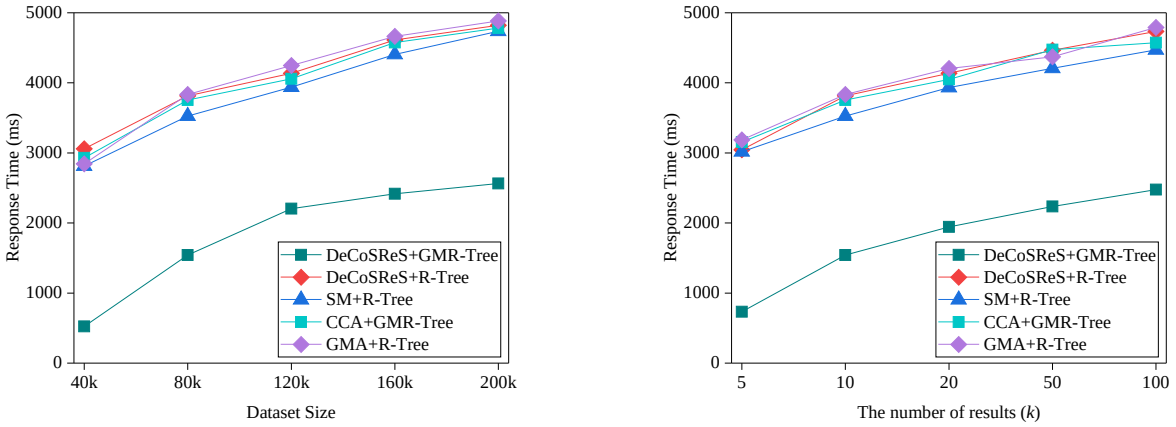
Input A GMR-Tree $\mathcal{G}$, a query $\mathcal{Q}$.
Output A results set $\mathcal{R}$.
Initializing: $\mathcal{R} \leftarrow \emptyset$;
Initializing: a priority queue $\mathcal{L} \leftarrow \emptyset$;
Initializing: an integer $\alpha \leftarrow 0$;
$\mathcal{L}.Enqueue(\mathcal{G}.Root, 0)$
for each element $M_{\mathcal{T}}^{(i)} \in \mathcal{Q}.M_{\mathcal{T}}$ **do**
  $M_{\mathcal{T}}^{(i)} \leftarrow \mathcal{H}_{SIG}(M_{\mathcal{T}}^{(i)})$;
end for
while $\alpha < \mathcal{Q}.k$ **do**
  $PTR_o \leftarrow NearestNeighbor(\mathcal{Q}.\lambda, \mathcal{Q}.M_{\mathcal{T}}, \mathcal{L})$
  $o \leftarrow LoadObject(PTR_o)$;
  if $\mathcal{F}_{score}(\mathcal{Q}, o) > \mathcal{F}_{score}(\mathcal{Q}, o'), \forall o' \in \mathcal{O} \setminus \mathcal{R}$ **then**
    $\mathcal{R} \leftarrow AddObject(o)$;
    $\alpha \leftarrow \alpha + 1$;
  end if
end while
return $\mathcal{R}$;

Algorithm 2 NearestNeighbor($\mathcal{Q}.\lambda$, $\mathcal{Q}.M_{\mathcal{T}}$, $\mathcal{L}$**)**

Input A query $\mathcal{Q}$, a list $\mathcal{L}$.
Output A results set $\mathcal{R}$.
Initializing: a variable $\mathcal{E} \leftarrow \emptyset$;
while $\mathcal{L}.IsNotEmpty()$ **do**
  $\mathcal{E} \leftarrow \mathcal{L}.Dequeue()$;
  if $\mathcal{E}$ is a non-leaf node **then**
    for each $\langle MBR, SIG, PTR_N \rangle$ in $\mathcal{E}$ **do**
      if *SIG* matches $\mathcal{Q}.M_{\mathcal{T}}$ **then**
        $\mathcal{L}.Enqueue(LoadNode(PTR_N), Dst(\mathcal{Q}.\lambda, MBR))$;
      end if
    end for
  else if $\mathcal{E}$ is a leaf node **then**
    for each $\langle MBR, SIG, PTR_o \rangle$ in $\mathcal{E}$ **do**
      if *SIG* matches $\mathcal{Q}.M_{\mathcal{T}}$ **then**
        $\mathcal{L}.Enqueue(LoadNode(PTR_o), Dst(\mathcal{Q}.\lambda, MBR))$;
      end if
    end for
  else
    return $\mathcal{E}$;
  end if
end while

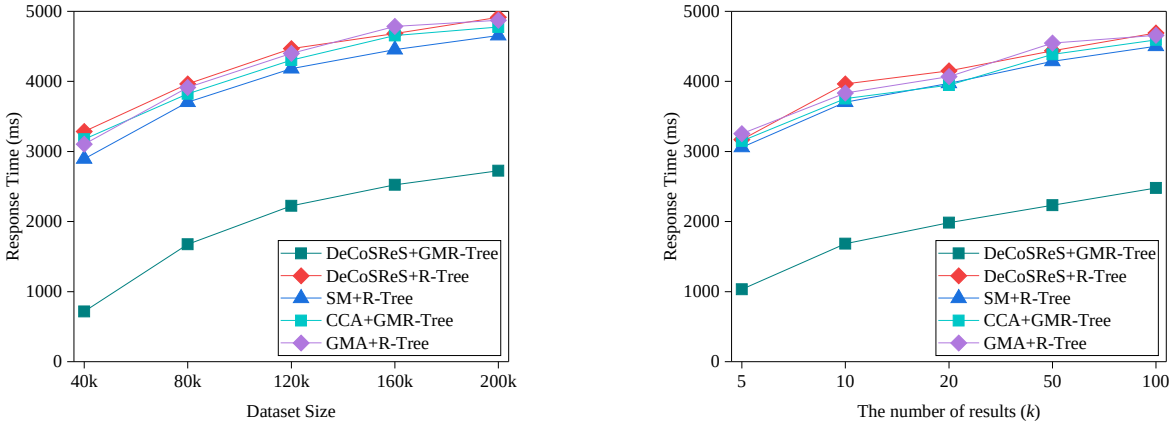
- **Flickr.** The real dataset Flickr includes over one million geo-tagged images that are crawled from Flickr (http://www.flickr.com/), a popular web site for users to share and embed personal photographs. To evaluate the scalability of our proposed algorithm, The dataset size varies from $40k$ to $200k$. The spatial locations of Flickr is obtained from the US Board on Geographic Names (http://geonames.usgs.gov).
- **ImageNet.** The synthetic dataset ImageNet is generated by obtaining the spatial locations from corresponding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access



(a) Different size of dataset (b) Different number of results

FIGURE 8: Evaluation on Flickr dataset



(a) Different size of dataset (b) Different number of results

FIGURE 9: Evaluation on ImageNet dataset

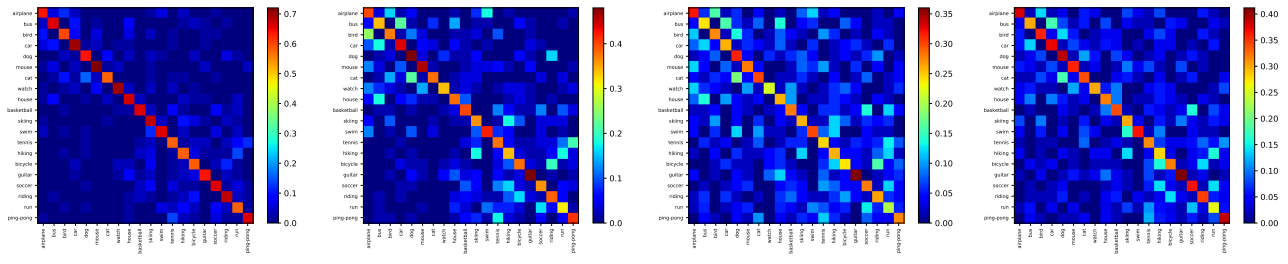
spatial dataset Rtree-Portal (http://www.rtreeportal.org) and randomly geo-tagging these objects with images in ImageNet (http://image-net.org/index). ImageNet is a famous image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. There are more than 100,000 synsets in WordNet, majority of them are nouns (80,000+). ImageNet provides on average 1000 images to illustrate each synset. Images of each concept are quality-controlled and human-annotated.

Some samples of Flickr and ImageNet dataset are shown in Fig. 7.

Workload. A workload for kNN geo-multimedia cross-modal query experiment includes 100 input queries. The query locations are randomly selected from the locations of the underlying objects. By default, the number of final results $k = 10$, and data number $N = 80k$. We use response time and precision to evaluate the performance of the algorithms. The size of dataset is set to $40k$, $80k$, $120k$, $160k$ and $200k$. The number of results k is set to 5, 10, 20, 50 and 100. Our experiments are run on a workstation with Intel(R) CPU Xeon 2.60GHz, 16GB memory and NVIDIA GeForce GTX 1080 GPU running Ubuntu 16.04 LTS Operation System. All algorithms in the experiments are implemented in Java and Python.

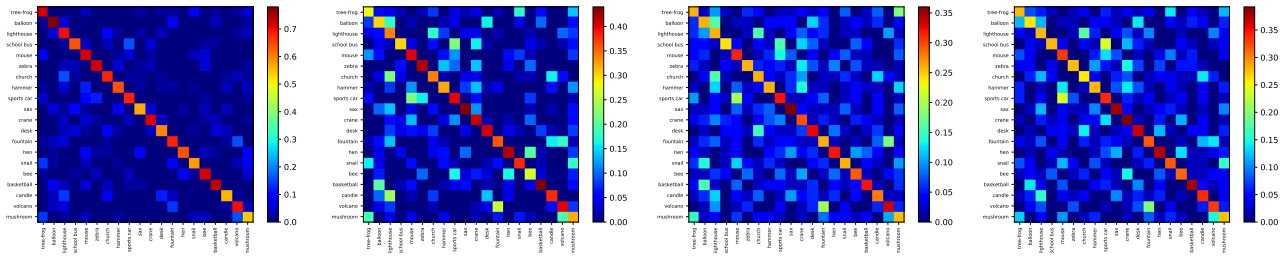
Baseline. To our best knowledge, this work is the first time to study the problem of kNN geo-multimedia cross-modal query. That means there is no existing approach for this problem. We devise four baseline methods, i.e., **DeCoSReS+R-Tree** and Semantic Matching [63]+R-Tree (**SM+R-Tree**), Canonical Correlation Analysis [67]+R-tree (**CCA+R-Tree**), and Generalized Multiview Analysis [98]+R-Tree (**GMA+R-Tree**), briefly introduced as fol-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access



(a) (b) (c) (d)

FIGURE 10: Cofusion matrices of classification precision on Flickr dataset. (a) DeCoSReS+GMR-Tree. (b) SM+R-Tree. (c) CCA+R-Tree. (d) GMA+R-Tree.



(a) (b) (c) (d)

FIGURE 11: Cofusion matrices of classification precision on ImageNet dataset. (a) DeCoSReS+GMR-Tree. (b) SM+R-Tree. (c) CCA+R-Tree. (d) GMA+R-Tree.

lows:

- **DeCoSReS+R-Tree**, the combination of the proposed deep learning based cross-modal retrieval method and R-Tree.
- **SM+R-Tree**, the combination of Semantic Matching and R-Tree. Semantic Matching model the semantic correlations between multi-modal data by learning a common semantic space.
- **CCA+R-Tree**, the combination of Canonical Correlation Analysis and R-Tree. Canonical Correlation Analysis aims to generate a common space by linear transformations to measure the correlations of multi-modal data.
- **GMA+R-Tree**, the combination of Generalized Multiview Analysis and R-Tree. Generalized Multiview Analysis uses labels of multi-modal data to learn the maps from multi-modal spaces to a common space. It is a kernelizable extension of CCA.

The feature representation technique used in these baselines is BoW model (BoVW for image), and the spatial area of geo-multimedia dataset is partitioned by R-Tree.

B. RESULTS OF EXPERIMENTS

1) Evaluation on Flickr Dataset

Evaluation on different size of dataset. We evaluate the performance of our approach DeCoSReS+GMR-Tree and four baselines, i.e., DeCoSReS+R-Tree, SM+R-Tree, CCA+R-Tree and GMA+R-Tree with the increment of dataset size. Fig. 8(a) shows how the variations of dataset size affect the search performance. With the increasing of dataset size, the response time of all these methods increase gradually. Not surprisingly, the proposed approach has the smallest response time due to the application of the proposed hybrid indexing structure GMR-Tree, which can speed up the spatial search markedly. It increases obviously and slow down when the dataset size is larger than $120k$. The efficiency of SM+R-Tree is a bit higher than DeCoSReS+R-Tree, which is showing a rise trend of volatility between $50k$ and $200k$. And at last, the response time of these two baselines are nearly 5000ms. The efficiency of CCA+R-Tree and GMA+R-Tree are similar to DeCoSReS+R-Tree. The response time of them rise with slight fluctuations and nearly 4950ms when the dataset size increases to $200k$, which is much higher than DeCoSReS+GMR-Tree. This verifies that the combination of semantic representation signature technique and MBR technique can outperform R-Tree for the task of geo-multimedia cross-modal retrieval.

Evaluation on different number of results k. We evaluate the performance of DeCoSReS+GMR-Tree, DeCoSReS+R-Tree, SM+R-Tree, CCA+R-Tree and GMA+R-Tree with the increasing of number of results k, as illustrated in Fig. 8(b). In this evaluation, we increase k from 5 to 100. Clearly, the response time of DeCoSReS+GMR-Tree is going up with the rising of k. When $k = 5$, the response time is smaller than 1000ms, and it increases step by step in the interval of $[10, 100]$. By contrast, the efficiency of other four approaches are much lower than the proposed method. Likewise, the response time of them climb step by step. Similar to the situation shown in Fig. 8(a), the performance of DeCoSReS+R-Tree, SM+R-Tree, CCA+R-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2940055, IEEE Access

Tree and GMA+R-Tree are similar, which are much lower than DeCoSReS+GMR-Tree.

2) Evaluation on ImageNet Dataset

Evaluation on different size of dataset. Fig. 9(a) illustrates the comparison of DeCoSReS+GMR-Tree, DeCoSReS+R-Tree, SM+R-Tree, CCA+R-Tree and GMA+R-Tree on the synthetic dataset ImageNet under the variations of dataset size. Obviously, the performances of these methods decrease step by step with the increasing of dataset size. By comparison, the proposed method defeats the opponents by an obvious superiority due to the benefit from GMR-Tree. When the dataset size is smaller than $100k$, the response time of it is less than 2000ms. On the other hand, the time cost of other four approaches are very close. They increase faster in the interval of $[50k, 100k]$. After that, the growth of them slow down. Like the comparison in the Flickr dataset, the search efficiency of R-Tree based methods cannot outperform the GMR-Tree based method.

Evaluation on different number of results k. Fig. 9(b) shows the evaluation of efficiency of DeCoSReS+GMR-Tree and other four opponents with the increment of number of results k. Similar to the situations on Flickr dataset, the efficiency of DeCoSReS+GMR-Tree slows down bit by bit with k increasing from 10 to 100. However, it is still the best approach among them due to the usage of GMR-Tree. The response time of other four algorithms are much higher than the proposed approaches. Like the evaluations above, the trends of DeCoSReS+R-Tree, SM+R-Tree, CCA+R-Tree and GMA+R-Tree are still similar since the same spatial search technique is employed. Specifically, they rise with slight fluctuations. At $k = 5$, they are nearly 3000ms. When $k = 100$, they increase to 4600ms around.

3) Evaluation on cross-modal retrieval precision

Evaluation on Flickr Dataset. Fig. 10 demonstrates that the confusion matrices of cross-modal retrieval on Flickr dataset by DeCoSReS+GMR-Tree, SM+R-Tree, CCA+R-Tree and GMA+R-Tree. The techniques of semantic representation space construction are different, which is the main factor affecting the retrieval precision. Specifically, the proposed method DeCoSReS+GMR-Tree employs AlexNet and LDA model for cross-modal feature representation as discussed in Section V, which has the best performance for the retrieval. The opponent SM+R-Tree uses SITF and BoVW to extract visual features in a traditional manner. Obviously, precision of it is lower than DeCoSReS+GMR-Tree. On the other hand, SM+R-Tree is a little bit better CCA+R-Tree and GMA+R-Tree due to the SM technique can represent multimodal semantic concepts precisely. However, all of these three methods are based on SIFT features that cannot represent the semantic correlations between different modalities, which is illustrated clearly by the comparison.

Evaluation on ImageNet Dataset. We compare the cross-modal classification precision of DeCoSReS+GMR-Tree with other three approaches on ImageNet dataset, shown as in Fig. 11. Similar to the evaluation on Flickr, the performance of our method is better obviously, which is benefit from the deep CNN based semantic representation space technique. For some classes, e.g. balloon, zebra and basketball, the precision of DeCoSReS+GMR-Tree is nearly 76%. On the other hand, SM+R-Tree, CCA+R-Tree and GMA+R-Tree cannot achieve such high precision.

VIII. CONCLUSION

In this paper, we propose a novel problem named kNN geo-multimedia cross-modal retrieval. It aims to return k nearest geo-multimedia objects that are highly similar to the query in the aspect of semantics. For the first time, we propose the definition of geo-multimedia object and kNN geo-multimedia cross-modal query, as well as the notion of cross-modal semantic representation space. To overcome this challenge, a novel framework of geo-multimedia cross-modal retrieval is proposed, which includes multi-modal feature extraction, cross-modal semantic space mapping, geo-multimedia spatial index and cross-modal semantic similarity measurement. To address the ticklish problem of semantic gap between different modalities, we present an approach called cross-modal semantic matching and an implementation via deep learning techniques to construct a common semantic representation space for multi-modal data. To speed up the geo-multimedia search, we propose a novel hybrid index structure, named GMR-Tree, which is a combination of R-Tree and signature files that are generated from the semantic representations of geo-multimedia objects. Based on it, we design an efficient kNN search algorithm named kGMCMS to support efficient geo-multimedia cross-modal retrieval. The experimental results show that our approach outperforms the-state-of-the-art methods.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61702560, 61472450, 61972203), the Key Research Program of Hunan Province(2016JC2018), project (2018JJ3691) of Science and Technology Plan of Hunan Province, and the Research and Innovation Project of Central South University Graduate Students(2018zzts177).

REFERENCES

[1] Wang, Y., Lin, X., Wu, L., & Zhang, W. (2015, October). Effective multi-query expansions: Robust landmark retrieval. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 79-88). ACM.

[2] Wang, Y., Lin, X., Wu, L., Zhang, W., & Zhang, Q. (2014, November). Exploiting correlation consensus: Towards subspace clustering for multi-modal data. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 981-984). ACM.

[3] Wang, Y., Wu, L., Lin, X., & Gao, J. (2018). Multiview spectral clustering via structured low-rank matrix factorization. IEEE transactions on neural networks and learning systems, (99), 1-11.

[4] Zhang, C., Chen, R., Zhu, L., Liu, A., Lin, Y., & Huang, F. (2018). Hierarchical information quadtree: efficient spatial temporal image search for multimedia stream. Multimedia Tools and Applications, 1-23.

[5] Wang, Y., Wenjie, Z., Wu, L., Lin, X., Fang, M., & Pan, S. (2016, January). Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In IJCAI International Joint Conference on Artificial Intelligence.

[6] Wang, Y., & Wu, L. (2018). Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. Neural Networks, 103, 1-8.

[7] Wang, Y., Lin, X., Zhang, Q., & Wu, L. (2014, May). Shifting hypergraphs by probabilistic voting. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 234-246). Springer, Cham.

[8] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

[9] Vukotić, V., Raymond, C., & Gravier, G. (2016, October). Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. In Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion (pp. 37-44). ACM.

[10] Cao, Y., Long, M., Wang, J., & Liu, S. (2017, February). Collective deep quantization for efficient cross-modal retrieval. In Thirty-First AAAI Conference on Artificial Intelligence.

[11] Wang, Y., Lin, X., Wu, L., Zhang, W., Zhang, Q., & Huang, X. (2015). Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Transactions on Image Processing, 24(11), 3939-3949.

[12] Wang, Y., Lin, X., Wu, L., Zhang, W., & Zhang, Q. (2015, August). Lbmch: Learning bridging mapping for cross-modal hashing. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 999-1002). ACM.

[13] Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. Multimedia systems, 16(6), 345-379.

[14] McNamara, Q., De La Vega, A., & Yarkoni, T. (2017, August). Developing a comprehensive framework for multimodal feature extraction. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1567-1574). ACM.

[15] Wang, Y., Lin, X., & Zhang, Q. (2013, October). Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (pp. 805-810). ACM.

[16] Wang, Y., Zhang, W., Wu, L., Lin, X., & Zhao, X. (2015). Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. IEEE transactions on neural networks and learning systems, 28(1), 57-70.

[17] Wu, L., Huang, X., Zhang, C., Shepherd, J., & Wang, Y. (2015). An efficient framework of Bregman divergence optimization for co-ranking images and tags in a heterogeneous network. Multimedia Tools and Applications, 74(15), 5635-5660.

[18] Bianchi-Berthouze, N. (2003). K-DIME: an affective image filtering system. IEEE MultiMedia, 10(3), 103-106.

[19] Li, Y., Zhang, Y., Tao, C., & Zhu, H. (2016). Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion. Remote Sensing, 8(9), 709.

[20] Chen, J., Wang, Y., Luo, L., Yu, J. G., & Ma, J. (2016). Image retrieval based on image-to-class similarity. Pattern Recognition Letters, 83, 379-387.

[21] Lowe, D. G. (1999, September). Object recognition from local scale-invariant features. In iccv (Vol. 99, No. 2, pp. 1150-1157).

[22] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[23] Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. In null (p. 1470). IEEE.

[24] Wu, L., Wang, Y., Gao, J., & Li, X. (2018). Deep adaptive feature embedding with local sample distributions for person re-identification. Pattern Recognition, 73, 275-288.

[25] Wu, L., Wang, Y., Li, X., & Gao, J. (2018). Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE transactions on cybernetics, (99), 1-12.

[26] LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010, May). Convolutional networks and applications in vision. In Proceedings of 2010 IEEE International Symposium on Circuits and Systems (pp. 253-256). IEEE.

[27] Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352-2449.

[28] Wang, Y., Lin, X., Wu, L., & Zhang, W. (2017). Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. IEEE Transactions on Image Processing, 26(3), 1393-1404.

[29] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[30] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[31] Hoffer, E., & Ailon, N. (2015, October). Deep metric learning using triplet network. In International Workshop on Similarity-Based Pattern Recognition (pp. 84-92). Springer, Cham.

[32] Huang, F., Zhang, X., Li, Z., Mei, T., He, Y., & Zhao, Z. (2017, October). Learning social image embedding with deep multimodal attention networks. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017 (pp. 460-468). ACM.

[33] Melekhov, I., Kannala, J., & Rahtu, E. (2016, December). Siamese network features for image matching. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 378-383). IEEE.

[34] Guo, J., Yue, B., Xu, G., Yang, Z., & Wei, J. M. (2017, April). An enhanced convolutional neural network model for answer selection. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 789-790). International World Wide Web Conferences Steering Committee.

[35] Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2015). Reasoning about entailment with neural attention. arXiv preprint arXiv:1509.06664.

[36] Wan, S., Lan, Y., Guo, J., Xu, J., Pang, L., & Cheng, X. (2016, March). A deep architecture for semantic matching with multiple positional sentence representations. In Thirtieth AAAI Conference on Artificial Intelligence.

[37] He, H., Gimpel, K., & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1576-1586).

[38] Yang, Y., Yuan, S., Cer, D., Kong, S. Y., Constant, N., Pilar, P., ... & Kurzweil, R. (2018). Learning semantic textual similarity from conversations. arXiv preprint arXiv:1804.07754.

[39] Wang, Z., Mi, H., & Ittycheriah, A. (2016). Sentence similarity learning by lexical decomposition and composition. arXiv preprint arXiv:1602.07019.

[40] Liu, K., Zhang, L., & Sun, Y. (2015, January). Text Retrieval analysis based on Deep Learning. In 2015 International Symposium on Computers & Informatics. Atlantis Press.

[41] Zhao, P., Kuang, X., Sheng, V. S., Xu, J., Wu, J., & Cui, Z. (2015, April). Scalable Top-k Spatial Image Search on Road Networks. In International Conference on Database Systems for Advanced Applications (pp. 379-396). Springer, Cham.

[42] Hariharan, R., Hore, B., Li, C., & Mehrotra, S. (2007, July). Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems. In 19th International Conference on Scientific and Statistical Database Management (SSDBM 2007) (pp. 16-16). IEEE.

[43] Cary, A., Wolfson, O., & Rishe, N. (2010, June). Efficient and scalable method for processing top-k spatial boolean queries. In International Conference on Scientific and Statistical Database Management (pp. 87-95). Springer, Berlin, Heidelberg.

[44] Cao, X., Cong, G., & Jensen, C. S. (2010). Retrieving top-k prestige-based relevant spatial web objects. Proceedings of the VLDB Endowment, 3(1-2), 373-384.

[45] Long, J., Zhu, L., Zhang, C., Yang, Z., Lin, Y., & Chen, R. (2018). Efficient interactive search for geo-tagged multimedia data. Multimedia Tools and Applications, 1-30.

[46] Zhang, D., Tan, K. L., & Tung, A. K. (2013, March). Scalable top-k spatial keyword search. In Proceedings of the 16th international conference on extending database technology (pp. 359-370). ACM.

[47] Zhang, C., Zhang, Y., Zhang, W., & Lin, X. (2016). Inverted linear quadtree: Efficient top k spatial keyword search. IEEE Transactions on Knowledge and Data Engineering, 28(7), 1706-1721.

[48] Zhang, C., Zhang, Y., Zhang, W., & Lin, X. (2016). Inverted linear quadtree: Efficient top k spatial keyword search. IEEE Transactions on Knowledge and Data Engineering, 28(7), 1706-1721.

[49] Li, Z., Lee, K. C., Zheng, B., Lee, W. C., Lee, D., & Wang, X. (2010). Ir-tree: An efficient index for geographic document search. IEEE Transactions on Knowledge and Data Engineering, 23(4), 585-599.

[50] Cong, G., Jensen, C. S., & Wu, D. (2009). Efficient retrieval of the top-k most relevant spatial web objects. Proceedings of the VLDB Endowment, 2(1), 337-348.

[51] Wu, D., Yiu, M. L., Cong, G., & Jensen, C. S. (2011). Joint top-k spatial keyword query processing. IEEE Transactions on Knowledge and Data Engineering, 24(10), 1889-1903.

[52] Wu, L., & Wang, Y. (2017). Robust hashing for multi-view data: Jointly learning low-rank kernelized similarity consensus and hash functions. Image and Vision Computing, 57, 58-66.

[53] Wu, L., Wang, Y., & Shepherd, J. (2013, October). Efficient image and tag co-ranking: a bregman divergence optimization method. In Proceedings of the 21st ACM international conference on Multimedia (pp. 593-596). ACM.

[54] Laenen, K., Zoghbi, S., & Moens, M. F. (2018, February). Web search of fashion items with multimodal querying. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 342-350). ACM.

[55] Yu, J., Yang, X., Gao, F., & Tao, D. (2016). Deep multimodal distance metric learning using click constraints for image ranking. IEEE transactions on cybernetics, 47(12), 4014-4024.

[56] Li, J., Wu, Y., Zhao, J., & Lu, K. (2016). Low-rank discriminant embedding for multiview learning. IEEE transactions on cybernetics, 47(11), 3516-3529.

[57] Jin, L., Li, K., Hu, H., Qi, G. J., & Tang, J. (2017). Semantic neighbor graph hashing for multimodal retrieval. IEEE Transactions on Image Processing, 27(3), 1405-1417.

[58] Rafailidis, D., Manolopoulou, S., & Daras, P. (2013). A unified framework for multimodal retrieval. Pattern Recognition, 46(12), 3358-3370.

[59] Moon, S., Kim, S., & Wang, H. (2014). Multimodal transfer deep learning with applications in audio-visual recognition. arXiv preprint arXiv:1412.3121.

[60] Dang-Nguyen, D. T., Piras, L., Giacinto, G., Boato, G., & Natale, F. G. D. (2017). Multimodal retrieval with diversification and relevance feedback for tourist attraction images. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 13(4), 49.

[61] Wang, X. J., Ma, W. Y., Zhang, L., & Li, X. (2005, November). Multi-graph enabled active learning for multimodal web image retrieval. In Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval (pp. 65-72). ACM.

[62] Min, W., Jiang, S., Sang, J., Wang, H., Liu, X., & Herranz, L. (2016). Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. IEEE Transactions on Multimedia, 19(5), 1100-1113.

[63] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010, October). A new approach to cross-modal multimedia retrieval. In Proceedings of the 18th ACM international conference on Multimedia (pp. 251-260). ACM.

[64] Bredin, H., & Chollet, G. (2007, April). Audio-visual speech synchrony measure for talking-face identity verification. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 2, pp. II-233). IEEE.

[65] Zhai, X., Peng, Y., & Xiao, J. (2012, March). Cross-modality correlation propagation for cross-media retrieval. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2337-2340). IEEE.

[66] Rasiwasia, N., Mahajan, D., Mahadevan, V., & Aggarwal, G. (2014, April). Cluster canonical correlation analysis. In Artificial Intelligence and Statistics (pp. 823-831).

[67] Hardoon, D. R. , Szedmak, S. , & Shawe-Taylor, J. . (2004). Canonical correlation analysis: an overview with application to learning methods. Neural Computation, 16(12), 2639-2664.

[68] Jia, Y., Salzmann, M., & Darrell, T. (2011, November). Learning cross-modality similarity for multinomial data. In 2011 International Conference on Computer Vision (pp. 2407-2414). IEEE.

[69] Chu, L., Zhang, Y., Li, G., Wang, S., Zhang, W., & Huang, Q. (2014). Effective multimodality fusion framework for cross-media topic detection. IEEE Transactions on Circuits and Systems for Video Technology, 26(3), 556-569.

[70] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

[71] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In Advances in neural information processing systems (pp. 2121-2129).

[72] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 689-696).

[73] Wu, L., Wang, Y., Ge, Z., Hu, Q., & Li, X. (2018). Structured deep hashing with convolutional neural networks for fast person re-identification. Computer Vision and Image Understanding, 167, 63-73.

[74] Wu, L., Wang, Y., Li, X., & Gao, J. (2018). What-and-where to match: deep spatially multiplicative integration networks for person re-identification. Pattern Recognition, 76, 727-738.

[75] Fu, R., Li, B., Gao, Y., & Wang, P. (2016, October). Content-based image retrieval based on CNN and SVM. In 2016 2nd IEEE International Conference on Computer and Communications (ICCC) (pp. 638-642). IEEE.

[76] Tolias, G., Avrithis, Y., & Jégou, H. (2013). To aggregate or not to aggregate: Selective match kernels for image search. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1401-1408).

[77] Tolias, G., Avrithis, Y., & Jégou, H. (2016). Image search with selective match kernels: aggregation across single and multiple images. International Journal of Computer Vision, 116(3), 247-261.

[78] Zhou, D., Li, X., & Zhang, Y. J. (2016, September). A novel CNN-based match kernel for image retrieval. In 2016 IEEE International Conference on Image Processing (ICIP) (pp. 2445-2449). IEEE.

[79] Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., & Li, J. (2014, November). Deep learning for content-based image retrieval: A comprehensive study. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 157-166). ACM.

[80] Pei-Xia, S., Hui-Ting, L., & Tao, L. (2016, August). Learning discriminative CNN features and similarity metrics for image retrieval. In 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC) (pp. 1-5). IEEE.

[81] Zagoruyko, S., & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4353-4361).

[82] Guttman, A. (1984). R-trees: a dynamic index structure for spatial searching (Vol. 14, No. 2, pp. 47-57). ACM.

[83] Cao, X., Cong, G., Jensen, C. S., & Ooi, B. C. (2011, June). Collective spatial keyword querying. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (pp. 373-384). ACM.

[84] Li, G., Xu, J., & Feng, J. (2012, October). Keyword-based k-nearest neighbor search in spatial databases. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2144-2148). ACM.

[85] Rocha-Junior, J. B., Gkorgkas, O., Jonassen, S., & Nørvåg, K. (2011, August). Efficient processing of top-k spatial keyword queries. In International Symposium on Spatial and Temporal Databases (pp. 205-222). Springer, Berlin, Heidelberg.

[86] Zhang, D., Chan, C. Y., & Tan, K. L. (2014, July). Processing spatial keyword query as a top-k aggregation query. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (pp. 355-364). ACM.

[87] Zhang, D., Chee, Y. M., Mondal, A., Tung, A. K., & Kitsuregawa, M. (2009, March). Keyword search in spatial databases: Towards searching by document. In 2009 IEEE 25th International Conference on Data Engineering (pp. 688-699). IEEE.

[88] Zhang, D., Chee, Y. M., Mondal, A., Tung, A. K., & Kitsuregawa, M. (2009, March). Keyword search in spatial databases: Towards searching by document. In 2009 IEEE 25th International Conference on Data Engineering (pp. 688-699). IEEE.

[89] Long, C., Wong, R. C. W., Wang, K., & Fu, A. W. C. (2013, June). Collective spatial keyword queries: a distance owner-driven approach. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (pp. 689-700). ACM.

[90] Hotelling, H. (1992). Relations between two sets of variates. In Breakthroughs in statistics (pp. 162-190). Springer, New York, NY.

[91] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[92] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[93] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[94] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[95] Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2), 270-280.

[96] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[97] Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf.
[98] Sharma, A., Kumar, A., Daume, H., & Jacobs, D. W. (2012, June). Generalized multiview analysis: A discriminative latent space. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2160-2167). IEEE.