**Manuscript version:** Author’s Accepted Manuscript  
The version presented in WRAP is the author’s accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**  
http://wrap.warwick.ac.uk/132745

**How to cite:**  
Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**  
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher’s statement:**  
Please refer to the repository item page, publisher’s statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.
An Efficient Approach for Geo-Multimedia Cross-Modal Retrieval

LEI ZHU1, JUN LONG1, CHENGYUAN ZHANG1, WEIREN YU2,3, XINPAN YUAN4, AND LONGZHI SUN1

1School of Computer Science and Engineering, Central South University, Changsha 410083, China
2School of Engineering and Applied Science, Aston University, Birmingham, U.K.
3Nanjing University of Science and Technology, China
4School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China

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

This work was supported in part by the National Natural Science Foundation of China under Grant 61702560, Grant 61472450, and Grant 61972203, in part by the Key Research Program of Hunan Province under Grant 2016JC2018, in part by the Project of Science and Technology Plan of Hunan Province under Grant 2018JJ3691, and in part by the Research and Innovation Project of Central South University Graduate Students under Grant 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 $k$NN 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 $k$NN search algorithm called $k$GMCMS 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, $k$NN 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\textsuperscript{1} has 1.15 billion users registered and the total number of images uploaded is 250 billion since its establishment. Twitter\textsuperscript{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\textsuperscript{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\textsuperscript{4}, which is the most popular photo shared web site and

\footnotesize\textsuperscript{1}https://facebook.com/
\footnotesize\textsuperscript{2}http://www.twitter.com/
\footnotesize\textsuperscript{3}https://weibo.com/
\footnotesize\textsuperscript{4}https://www.flickr.com/
it had a total of 87 million registered users. For the video 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 data has different modalities, 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\(^18\), IRMFRCAFM\(^19\) 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 is an example of traditional cross-modal retrieval.

Example 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

\(^5\)https://www.youtube.com/
\(^6\)http://www.iqiyi.com/
\(^7\)https://www.wikipedia.org/
spouts of whales, etc., which are similar to the geysers in the aspect of visual 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], [47], IR-tree [48] and its variations [49], WIR-tree [50], 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 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 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 \( k \) GMCMS is developed to implement \( k \) NN 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 \( k \) NN 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 \( k \) GMCMS 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 \( k \) NN 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 \( k \) GMCMS is developed to support geo-multimedia cross-modal query. Our experimental results are presented in Section VII, and finally we draw the conclusion 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.

1) **MULTI-MODAL RETRIEVAL**

Multi-modal retrieval [51] aims to search multimedia data [52] with multiple modalities. Laenen et al. [53] 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 ranking problem, Yu et al. [54] 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 [55]. Jin et al. [56] 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. [57] 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. [58] 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. [59] 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. [60] presented a multi-modal web image retrieval technique to leverage the heterogeneous data on the web to improve retrieval precision. In this solution, three graphs, i.e., Content-Graph, Text-Graph and Link-Graph which are constructed on visual content features,
textual annotations and hyperlinks respectively, provide complementary information on the images. To solve the problem of recipe-oriented image-ingredient correlation learning, Min et al. [61] proposed a multi-modal multitask deep belief network (M2TDBN) to learn joint image-ingredient representation regularized by different attributes.

2) 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 [62]. To exploit the correlation between multiple modalities, Bredin and Chollet [63] utilized canonical correlation analysis (CCA) [66] 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. [64] 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. [65] 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. [62] 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. [67] presented a novel Markov random field based model which learns cross-modality similarity from a document corpus that has multinomial data. Chu et al. [68] 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 solved by new models via deep neural networks [69]–[73]. 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. [74] 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 [75], [76] methods, Zhou et al. [77] 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. [78] 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. Pei-Xia et al. [79] 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 and Komodakis [80] 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 [81], which uses minimum bounding area (MBR) to partition the geographical space. Cao et al. [82] 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. [49] presented a new indexing framework that integrates the inverted file for text retrieval and the R-tree for spatial proximity querying. Li et al. [83] 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, Rocha-Junior et al. [84] 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. [48] 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. [85] 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. [86] introduced a new spatial keyword query problem called $m$-closest keywords ($m$CK) 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 [86]. Moreover, They exploited a priori-based search strategy to effectively reduce the search space. For collective spatial keyword query problem, Long et al. [87] 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 investigate 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 $k$NN 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 $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$, 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 $\mathcal{M} = \{T, I\}$, where $T$ represents text modality and $I$ represents image modality. If a geo-multimedia object contains a text, it is denoted as $o.M_T$. Similarly, If an object contains an image, it is denoted as $o.M_I$. $M_T$ and $M_I$ denote the feature vector generated by a text and an image respectively. Let $\mathcal{S}_T$ and $\mathcal{S}_I$ be the feature spaces of text and image, $\forall o_i \in \mathcal{O}$, if $o_i$ contains a text, then $o_i.M_T \in \mathcal{S}_T$. If $o_i$ contains an image, then $o_i.M_I \in \mathcal{S}_I$.

Based on the definition of geo-multimedia objects, we define the $k$NN 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 (Cross-Modal Query): Given a multimedia objects database $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$, 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) $Q_{\mathcal{T}T2I}$ is defined as a text query which aims to search out the most relevant multimedia object $o \in \mathcal{O}$ contains an image, and $Q_{\mathcal{T}2I}.M_T \in \mathcal{S}_T$, $o.M_I \in \mathcal{S}_I$. (2) $Q_{\mathcal{T}2I}^{\mathcal{I}T}$ is defined as a query image which aims to search out the most relevant multimedia object $o \in \mathcal{O}$ contains a text, and $Q_{\mathcal{T}2I}.M_I \in \mathcal{S}_I$, $o.M_T \in \mathcal{S}_T$.

Definition 3 (kNN Geo-Multimedia Cross-Modal Query): Given a geo-multimedia objects database $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$, a kNN Geo-Multimedia Cross-Modal Query $Q^k = (\lambda, M)$ 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 $Q^k_{\mathcal{T}T2I}$ and $Q^k_{\mathcal{T}2I}^{\mathcal{I}T}$, which are named kNN geo-multimedia text to image query ($kT2IQ$) and kNN geo-multimedia image to text query ($kI2TQ$) respectively. In more detail, $Q^k_{\mathcal{T}T2I}$ aims to return $k$ nearest geo-multimedia objects which contain images that are highly relevant to the query text, and $Q^k_{\mathcal{T}2I}^{\mathcal{I}T}$ 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 $Q_{\mathcal{T}T2I}$, the result is $k$ geo-multimedia objects $R_{\mathcal{T}T2I}$ which are ranked by the a score function $F_{\text{score}}(Q^k_{\mathcal{T}T2I}, o)$, i.e.,

$$R_{\mathcal{T}T2I} = \{o|o \in \mathcal{O}, o' \in \mathcal{O} \setminus R_{\mathcal{T}T2I}, F_{\text{score}}(Q^k_{\mathcal{T}T2I}, o) > F_{\text{score}}(Q^k_{\mathcal{T}T2I}, o')\},$$

$$R_{\mathcal{T}T2I} \subseteq \mathcal{O}, \ |R_{\mathcal{T}T2I}| = k$$

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

$$R_{\mathcal{I}2T} = \{o|o \in \mathcal{O}, o' \in \mathcal{O} \setminus R_{\mathcal{I}2T}, F_{\text{score}}(Q^k_{\mathcal{T}2I}^{\mathcal{I}T}, o) > F_{\text{score}}(Q^k_{\mathcal{T}2I}^{\mathcal{I}T}, o')\},$$

$$R_{\mathcal{I}2T} \subseteq \mathcal{O}, \ |R_{\mathcal{I}2T}| = k$$

and the score function is defined as follows:

$$F_{\text{score}}(Q, o) = \mu \text{Dst}(Q, o) + (1 - \mu)\text{Sim}(Q, o)$$

where $Q$ 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 $Q$ and $o$.

In this paper, we focus on the $kT2IQ$ query $Q^k_{\mathcal{T}T2I}$: 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 $Q^k_{\mathcal{T}T2I}$ as $Q$. 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 $\mathcal{O} = \{o_1, o_2, \ldots,$ $o_{|\mathcal{O}|}\}$ and a $kT2IQ$ query $Q$, $\forall o \in \mathcal{O}$, the spatial distance proximity is measured by the following function:

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

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

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

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

where the function max($\lambda$) is to return the maximum value of element in the set $\lambda$. 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 $\mathcal{O} = \{o_1, o_2, \ldots, o_{|\mathcal{O}|}\}$ and a $kT2IQ$ query $Q$, $\forall o \in \mathcal{O}$, the cross-modal semantic similarity is measured by cosine similarity measurement, as shown in the following equation:

$$\text{Sim}(Q, o) = \frac{\sum_{i \in \mathcal{O}.M_T} Q.M_T^{(i)} \ast o.M_I^{(i)}}{\sqrt{\sum_{i \in \mathcal{O}.M_T} (Q.M_T^{(i)})^2} \ast \sqrt{\sum_{i \in \mathcal{O}.M_I} (o.M_I^{(i)})^2}}$$

where $Q.M_T^{(i)}$ and $o.M_I^{(i)}$ represent $i$th feature element in representation vector $Q.M_T$ and $o.M_I$ 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 $Q.M_T$ and $o.M_I$ 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 $Q$ with a text and a geo-multimedia object $o$ with an image, the features spaces of them are denoted as $\mathcal{S}_T$ and $\mathcal{S}_I$ respectively, and $Q.M_T \in \mathcal{S}_T$, $o.M_I \in \mathcal{S}_I$, the mapping between $\mathcal{S}_T$ and $\mathcal{S}_I$ is represented as

$$\psi : \mathcal{S}_T \rightarrow \mathcal{S}_I$$

and the inverse mapping is represented as

$$\psi^{-1} : \mathcal{S}_I \rightarrow \mathcal{S}_T$$

Thus, the cross-modal text to image query can be denoted as $Q_{\mathcal{T}T2I} \iff \psi(Q.M_T)$. 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 $\mathbb{W}_T$ and $\mathbb{W}_I$ respectively, that is:

$$\Omega_T : \mathcal{S}_T \rightarrow \mathbb{W}_T$$

$$\Omega_I : \mathcal{S}_I \rightarrow \mathbb{W}_I$$
This idea is to use two intermediate representation spaces \( W \) isomorphic spaces that means there is a semantic correlation between these two and existing a mapping \( \Omega \).

Given a geo-multimedia database cross-modal text to image query in the following forms:

- Learn the intermediate representation spaces for implementing efficient cross-modal retrieval is to
- To overcome this challenge, we introduce a notion named
- 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 designed 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.

A. 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 \( \psi(\psi(I_1, I_2, ..., I_m); \theta) = \{t_1, t_2, ..., t_m\}, \quad \psi(T_1, T_2, ..., T_m); \psi) = \{r_1, r_2, ..., r_m\} \), where \( \theta \) and \( \psi \) 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 andTxtNet, which are explained minutely in Section V. Other techniques will be exploited in our future works.

B. 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 high-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 Cross-modal Semantic Matching (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.

C. 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 information 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.

D. SIMILARITY MEASUREMENT AND SEARCH

Based on GMR-Tree, we design a kNN geo-multimedia cross-modal search algorithm, called kGMCMS. The score function $f_{score}(Q, o) = \mu Dist(Q, o) + (1 - \mu)Sim(Q, o)$ defined in Section III is used to measure the similarity between the query $Q$ and the geo-multimedia object $o$ in both aspects of geo-location and semantic concept precisely.

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 [89] (LDA) to learn the representation space.

A. CROSS-MODAL SEMANTIC MATCHING

We use the method called cross-modal semantic matching (CoSMat) to construct CoSReS so that it provides a common semantic representation space for different modalities. This algorithm consists of two components, i.e., (1)CCA based Correlation Projection (CorrProj) and (2)logistic regression based Transformation (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 [88] (CCA) is a popular dimensionality reduction method. We use it to learn \( \gamma \)-dimensional subspaces \( \mathbb{W}_T' \subset \mathbb{S}_T \) and \( \mathbb{W}_I' \subset \mathbb{S}_I \) to find the correlations between these two subspaces. CCA method learns directions in text and image feature spaces, i.e., \( \Gamma_T \in \mathbb{S}_T \) and \( \Gamma_I \in \mathbb{S}_I \) along the directions of the data maximally correlated. That is, for feature vectors \( M_T \) and \( M_I \), measuring the maximum correlation:

\[
u = \Gamma_T^\top M_T, \quad \mathbf{v} = \Gamma_I^\top M_I,
\]

\[
\max \text{Corr}(u, v) = \frac{\Gamma_T^\top \Sigma_{II} \Gamma_I}{\sqrt{\Gamma_T^\top \Sigma_{TT} \Gamma_T} \sqrt{\Gamma_I^\top \Sigma_{II} \Gamma_I}}
\]

wherein \( \Sigma_{TT} \) and \( \Sigma_{II} \) are the empirical covariance matrices of space \( \mathbb{S}_T \) and \( \mathbb{S}_I \), i.e., \( \Sigma_{TT} = \text{Cov}(\mathbb{S}_T) \) and \( \Sigma_{II} = \text{Cov}(\mathbb{S}_I) \). \( \Gamma_{TT} \) is the empirical cross-covariance matrix of them, i.e., \( \Sigma_{TT} = \text{Cov}(\mathbb{S}_T, \mathbb{S}_I) \), and \( \Sigma_{II} = \Gamma_{II} \).

The first \( \gamma \) canonical components \( \{ \Gamma_T \}^\gamma \) and \( \{ \Gamma_I \}^\gamma \) represent a basis for projection \( \mathbb{S}_T \) and \( \mathbb{S}_I \) on \( \mathbb{W}_T \) and \( \mathbb{W}_I \). For each text \( M_T \) in space \( \mathbb{S}_T \), it can be mapped into the projection \( \Theta_T(M_T) \) onto \( \{ \Gamma_T \}^\gamma \). Likewise, for each image \( M_I \) in space \( \mathbb{S}_I \), it can be mapped into the projection \( \Theta_I(M_I) \) onto \( \{ \Gamma_I \}^\gamma \). Therefore, the method CorrProj can learn two projections \( \Theta_T(M_T) \) and \( \Theta_I(M_I) \) from \( \mathbb{S}_T \) and \( \mathbb{S}_I \), which can be used to define two \( \gamma \)-dimension subspaces for text and image, i.e.,

\[
\Theta_T : \mathbb{S}_T \rightarrow \mathbb{W}_T
\]

and,

\[
\Theta_I : \mathbb{S}_I \rightarrow \mathbb{W}_I
\]

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

LogsTran. The method aforementioned is to map feature spaces of text and image to maximally correlated subspaces \( \mathbb{W}_T \) and \( \mathbb{W}_I \). Then we use another method called LogsTran to find the correspondence between \( \mathbb{S}_T \) and \( \mathbb{S}_I \) 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 \( C = \{ c_1, c_2, \ldots, c_n \} \), such as “airplane,” “cat” or “house.” We utilize logistic regression to learn two transformation \( \mathcal{L}_T \) and \( \mathcal{L}_I \). \( \mathcal{L}_T \) transforms a text contained by a geo-multimedia object \( o \in \mathbb{S}_T \) into a vector of posterior probabilities \( P_{\mathcal{L}_T}(u_i|\mathcal{T}) \), in which \( \mathcal{T} = \{ u_1, u_2, \ldots, u_k \} \) is a set of classes. Likewise, \( \mathcal{L}_I \) transforms an image contained by a geo-multimedia object \( o \in \mathbb{S}_I \) into a vector of posterior probabilities \( P_{\mathcal{L}_I}(u_i|\mathcal{I}) \). The spaces \( \mathbb{R}_T \) and \( \mathbb{R}_I \) of these posterior probabilities vectors are referred to the semantic representation space of text and image respectively. Formally, they can be presented as follows:

\[
\mathcal{L}_T : \mathbb{S}_T \rightarrow \mathbb{R}_T
\]

\[
\mathcal{L}_I : \mathbb{S}_I \rightarrow \mathbb{R}_I
\]

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:

\[
P_{\mathcal{L}_T}(c_i|M_i; \varpi) = \frac{1}{\sum_{c_j} \exp(\varpi_{cj}^\top M_i)} \exp(\varpi_{ci}^\top M_i)
\]

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

According to the logistic regression, in semantic representation spaces \( \mathbb{R}_T \) and \( \mathbb{R}_I \), 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 \( \mathbb{R}_T \) and \( \mathbb{R}_I \) are isomorphic, and they can be regarded as the same, i.e., \( \mathbb{R}_T = \mathbb{R}_I \). Therefore, the cross-modal semantic representation space \( \mathcal{W} = \mathbb{R}_T = \mathbb{R}_I \).

The CoSMat method is a combination of CorrProj and LogsTran. In the first step, CorrProj is applied to learn two maximally correlated subspaces \( \mathbb{W}_T \) and \( \mathbb{W}_I \) based on feature spaces \( \mathbb{S}_T \) and \( \mathbb{S}_I \). Then LogsTran method is used to generate two transformations \( \mathcal{L}_T \) and \( \mathcal{L}_I \) to create the isomorphic semantic representation spaces \( \mathbb{R}_T \) and \( \mathbb{R}_I \). Thus, we can measure the semantic similarity of text and image in the CoSReS \( \mathcal{W} \), i.e., \( \text{Sim}(\xi_T, \xi_I) \), where \( \xi_T = \mathcal{L}_T(\Theta_T(S_T)) \), \( \xi_I = \mathcal{L}_I(\Theta_I(S_I)) \). It is an significant step of implementing kT1Q.
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 224 × 224 × 3 input image, which has 96 kernels of size 11 × 11 × 3. The second convolutional layer has 256 kernels of size 5 × 5 × 96. The third convolutional layer has 384 kernels of size 3 × 3 × 256. The fourth convolutional layer has 384 kernels of size 3 × 3 × 192. The fifth convolutional layer has 256 kernels with size of 3 × 3 × 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 several topics. Specifically, a text is modeled by a multinomial distribution over \( \kappa \) topics and each word in a text is generated by first sampling a topic from the text-specific topic distribution [89].

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 [90], GoogLeNet [91] and ResNet [92] for image, and RNN [93], BiLSTM [94], [95] 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 \( \mathfrak{W}_T \) and \( \mathfrak{W}_I \) are built by CorrProj from the textual and visual feature vectors. Then, two semantic mappings are learned from \( \mathfrak{W}_T \) and \( \mathfrak{W}_I \) by LogsTran. That means \( \Sigma_T \) and \( \Sigma_I \) 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 Geo-Multimedia 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 GMR-Tree and related query.

A. HYBRID INDEXING STRUCTURE

The proposed hybrid index is called GMR-Tree. It is a combination of an R-Tree [81] 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 \( \langle \text{MBR}, \text{SIG}, \text{PTR} \rangle \) contains three components. \( \text{MBR} \) is defined as in the R-Tree, which represents the geo-location in the form of minimum bounding area (MBR). \( \text{SIG} \) is a signature file generated from the geo-multimedia objects in this MBR. For the \( i \)th object \( o_i \) in MBR, its signature is denoted as \( S_i = \mathcal{H} \text{SIG}(o_i, M_I) \), wherein \( \mathcal{H} \text{SIG}(\cdot) \) is a hashing function which is used to generate a signature from the semantic representation vector. For a MBR, the signature \( \text{SIG} = S_1 \lor S_2 \lor \ldots \lor S_i \), wherein the operator \( \lor \) 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 \( \text{PTR} \), which refers to a subnode. Similarly, the leaf node in GMR-Tree is the form of \( \langle \text{MBR}, \text{SIG}, \text{PTR} \rangle \) but the pointer \( \text{PTR} \) 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 \( Q \) and a node \( N_i \), the signatures of \( Q \) and \( N_i \) are \( \text{SIG}_Q \) and \( \text{SIG}_{N_i} \) respectively. If \( \text{SIG}_Q = \text{SIG}_Q \lor \text{SIG}_G \), that means the query \( Q \) 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, \( Q \) 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 \( k \text{GMCMS} \) algorithm is demonstrated in Algorithm 1. Algorithm 2 is the GMR-Tree based nearest neighbor search algorithm that is used in \( k \text{GMCMS} \).
Algorithm 2 NearestNeighbor($Q, \lambda^*, Q.M_T, L$)

1: Input A query $Q$, a list $L$
2: Output A results set $R$
3: Initializing: a variable $E \leftarrow \emptyset$
4: while $L.IsNotEmpty()$ do
   5: $E \leftarrow L.Dequeue()$
   6: if $E$ is a non-leaf node then
      7: for each $(MBR, SIG, PTR_N)$ in $E$ do
         8: if SIG matches $Q.M_T$ then
            9: $L.Enqueue(LoadNode(PTR_N), Dst(Q, MBR))$
         end if
      end for
   else if $E$ is a leaf node then
      11: for each $(MBR, SIG, PTR_o)$ in $E$ do
         12: if SIG matches $Q.M_T$ then
            13: $L.Enqueue(LoadNode(PTR_o), Dst(Q, MBR))$
         end if
      end for
   else
      17: return $E$
   end if
end while

For Algorithm 1, in the first step, a priority queue $L$ is initialized as a empty set and an integer $\alpha$ which is used for counting during the search. $R$ is the set of results. First the algorithm puts the root node of GMR-Tree $G$ into $L$, and then generates the signature for query $Q$. In this process, each element of semantic representation vector $Q.M_T$ is reassigned by a hashing function $H(SIG)$ that converts the element of $Q.M_T$ into a hash code. After that, the search process is implemented by a While loop. During the process, the nearest neighbor $o$ of query $Q$ is found out and then the score of $o$ is calculated by score function $f_{score}(Q, o)$ 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 $E$ to store a tree node. $L$ will be checked circularly whether it is empty or not. If $L$ is not empty, the algorithm gets a node stored in $L$ by a $Dequeue(.)$ operation and put it into $E$. If this node is a non-leaf node, and exist an object whose SIG matches the query, then measures the distance between $Q$ and $MBR$ of $E$. It will be put into $L$ again. If $E$ is a leaf node, all objects in it will be checked and put the object which matches the query in to $L$.

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

- **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 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 [62]+R-Tree (SM+R-Tree), Canonical Correlation Analysis [66]+R-tree (CCA+R-Tree), and Generalized Multiview Analysis [96]+R-Tree (GMA+R-Tree), briefly introduced as follows:

- **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 Multi-view Analysis and R-Tree. Generalized Multi-view 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

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.

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

2) EVALUATION ON IMAGENET DATASET

a: 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.

b: 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

a: 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 SIFT 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.

b: 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 $k$NN 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 $k$NN 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 $k$NN 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.

REFERENCES


LEI ZHU was born in Changsha, China, in 1988. He received the M.Sc. degree from Central South University, China, in 2014, where he is currently pursuing the Ph.D. degree in computer science and technology with the School of Computer Science and Engineering. His current research interests include machine learning, deep learning, computer vision, and spatio-temporal data retrieval.
JUN LONG received the M.Sc. and Ph.D. degrees in computer science from Central South University, China, in 2003 and 2011, respectively, where he is currently a Professor with the School of Information Science and Engineering. His current research interests include network management, QoS guarantees, web service, and multimedia systems and retrieval.

CHENGYUAN ZHANG was born in Hunan, China. He received the B.S. degree from Sun Yat-sen University, in 2008, and the master’s and Ph.D. degrees in computer science from the University of New South Wales, in 2011 and 2015, respectively. He is currently a Lecturer with the School of Computer Science and Engineering, Central South University, China. His current research interests include information retrieval, query processing on spatial data, and multimedia data.

WEIREN YU received the Ph.D. degree from the School of Computer Science and Engineering, University of New South Wales. He was a Postdoctoral Researcher with Imperial College London. He is currently a Lecturer in computer science with Aston University, and an Honorary Fellow with Imperial College London. He has published more than 30 articles in DB and IR. His current research interests include web search and information retrieval, graph data management, and streams data mining. He received three Best Paper Awards, two CiSRA Best Paper Awards, One of the Best Papers of ICDE 2013, and the Best Student Paper Award. He has served on various editorial boards, and as a PC and an Active Reviewer of journals, including IEEE TKDE, VLDB J, IEEE TIFS, ACM TKDD, WWWJ, and Sensors, and conferences, including SIGIR, SIGMOD, VLDB, ICDE, EDBT, and CIKM.

XINPAN YUAN received the Ph.D. degree in information science and engineering from Central South University, in 2012. He is currently a Lecturer with the School of Computer, Hunan University of Technology, Zhuzhou, China. His current research interests include natural language processing and information retrieval.

LONGZHI SUN was born in Shandong, China, 1988. He received the M.Sc. degree in engineering from Central South University, China, in 2016, where he is currently pursuing the Ph.D. degree in computer science and technology. His current research interests include deep learning, big data, and intelligent software systems.