Article

Efficient Multimedia Similarity Measurement Using Similar Elements

Jun Long [1], Lei Zhu [1], Xinpan Yuan [2]*, Longzhi Sun [1]

[1] School of Computer Science and Engineering, Central South University, Changsha, 410083, PR China;
[2] School of Computer Science, Hunan University of Technology, Zhuzhou, 412007, PR China
* Correspondence: xpyuan@hut.edu.cn

Abstract: Online social networking techniques and large-scale multimedia retrieval are developing rapidly, which not only has brought great convenience to our daily life, but generated, collected, and stored large-scale multimedia data as well. This trend has put forward higher requirements and greater challenges on massive multimedia retrieval. In this paper, we investigate the problem of image similarity measurement, which is one of the key problems of multimedia retrieval. Firstly, the definition of similarity measurement of images and the related notions are proposed. Then, an efficient similarity measurement framework is proposed. Besides, we present a novel basic method of similarity measurement named SMIN. To improve the performance of similarity measurement, we carefully design a novel indexing structure called SMI Temp Index (SMII for short). Moreover, we establish an index of potential similar visual words off-line to solve to problem that the index cannot be reused. Experimental evaluations on two real image datasets demonstrate that the proposed approach outperforms state-of-the-arts.

Keywords: Image Similarity; SMI; SMI Temp Index; PSMI

1. Introduction

In the recent years, online social networking techniques and large-scale multimedia systems [1–6] are developing rapidly, which not only has brought great convenience to our daily life, but generated, collected, and stored large-scale multimedia data [7,8], such as text, image [9], audio, video [10]. For example, in China, Weibo (https://weibo.com/) has 376 million active users and more than 100 million micro-blogs containing short text, image, or short video are posted. The most famous social networking platform all over the world, Facebook (https://facebook.com/), reports 350 million images uploaded everyday in the end of November 2013. More than 400 million tweets with texts and images have been generated by 140 million users on Twitter (http://www.twitter.com/). Another type of common application is multimedia data sharing services. Flickr(https://www.flickr.com/) is one of the most famous photos sharing web site around the world. More than 3.5 million new images uploaded to this platform every day in March 2013. More than 14 million articles are clicked every day on Pinterest (https://www.pinterest.com/). More than 2 billion totally videos stored in YouTube (https://www.youtube.com/), and every minute there are 100 hours of videos which are uploaded to this service. The total watch time exceeded 42 billion minutes on IQIYI (http://www.iqiyi.com/), the most famous online video sharing service in China and number of independent users monthly is more than 230 million monthly. For audio sharing services, the total amount of audio in Himalaya (https://www.ximalaya.com/) had exceeded 15 million as of December 2015. Other web services like Wikipedia (https://en.wikipedia.org/), the largest and most popular free encyclopedia on the Internet, contains more than 40 million articles with pictures in 301 different languages. Other mobile applications such as WeChat, Instagram, etc, provide great convenience for us to share multimedia data.

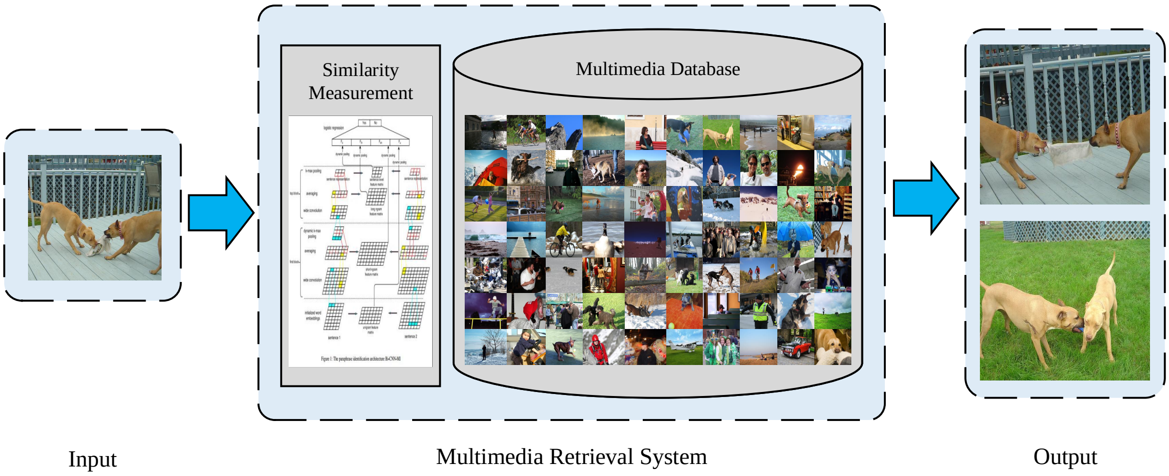


Figure 1. An example of multimedia retrieval via similarity measurement

Thanks to these current rich multimedia services and applications, multimedia techniques [11,12] is changing every aspect of our lives. On the other hand, the emergence of massive multimedia data [13] and applications puts forward greater challenges for techniques of information retrieval.

Motivation. Textual similarity measurement is a classical issue in the community of information retrieval and data mining. Lots of approaches have been proposed to improve the performance of similarity measurement. Guo et al [14] proposed to use vectors as basic elements, and the edit distance and Jaccard coefficient are used to calculate the sentence similarity. Li et al. [15] proposed the use of word vectors to represent the meaning of words, and considers the influence of multiple factors such as word meaning, word order and sentence length on the calculation of sentence similarity. Unlike the studies of textual similarity measurement, this work investigates the problem of image similarity measurement, which is a widely applied technique in lots of application scenarios, such as image retrieval [16–19], image near duplicate detection and matching [20,21]. There are two examples shown in Figure 1 and Figure 2 which can describe this problem clearly.

Example 1: In Figure 1, An user has a photo and she want to find out other pictures which are highly similar to it. She can submit an image query into the multimedia retrieval system. The system measures the visual similarity between this photo and the images in the database and after that a set of similar images is returned.

Example 2: Figure 2 demonstrates another application of image similarity measurement. An user want to measure similarity betweeen two pictures in a dataset quantitatively. She selects two pictures from the image dataset and input them into the similarity measurement system. According to image similarity measurement algorithm, the system will calculate the value of similarity between these images (e.g., 90%).

To improve the efficiency and accuracy of image similarity measurement, we present the definition of similarity measurement of images and the relevant notions. An efficient image similarity measurement framework is proposed, in which a coupled CNNs model is used to learn the deep visual feature representations. Compared to the traditional manner (e.g., SIFT), the deep CNN based method can capture more high level semantic features. Besides, we introduce the measurement of similar visual words named SMI Naive (SMIN for short) which is the basic method for similarity measurement, and then propose the SMIN algorithm. After that, to optimize this method, we design a novel indexing structure named SMI Temp Index to reduce the time complexity of calculation. In addition, another technique named index of potential similar visual words is proposed to solve the problem that the index cannot be reused. We could search for the index to perform the measurement of similar visual words without having to repeatedly create a temporary index.

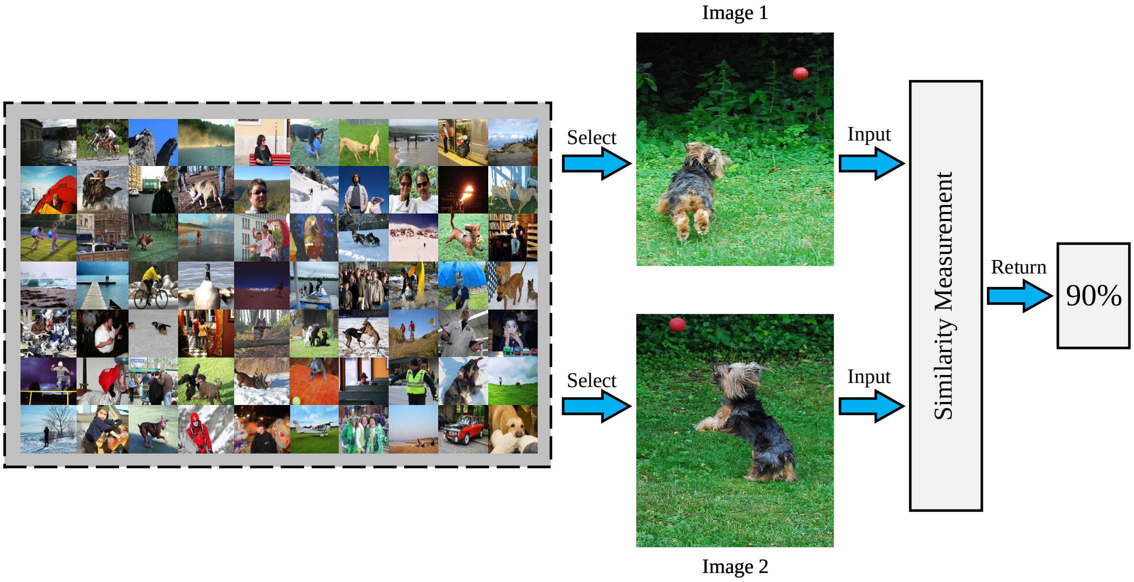


Figure 2. An example of multimedia retrieval via similarity measurement

Contributions. Our main contributions can be summarized as follows:

- The definition of similarity measurement of images and the related conceptions are introduced. Besides, the image similarity measurement function are designed.
- An efficient image similarity measurement framework is designed, which is a combination of a coupled CNNs module, BoVW module and similarity measurement module. Besides, the basic method of image similarity measurement is proposed, which is called SMI Naive (SMIN for short). To improve the performance of similarity measurement, we design two indexing techniques named SMI Temp Index (SMII for short) and Index of Potential Similar Visual Words (PSMI for short).
- Extensive experiments are conducted on two real image datasets. Experimental results demonstrate that our solution outperforms the state-of-the-art method.

Roadmap. In the remainder of this paper, Section 2 presents the related works about image similarity measurement and image retrieval. In Section 3, the definition of image similarity measurement and related conceptions are proposed. We present the basic similarity measurement method named SMIN and two improved indexing techniques and algorithms in Section 4. Our experimental results are presented in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

In this section, we review the related works of image similarity measurement and image retrieval, which are relevant to this study.

Image Similarity Measurement. In recent years, image similarity measurement has become a hot issue in the community of multimedia system [22,23] and information retrieval since the massive image data can be accessed in the Internet. On the other hand, like textual similarity measurement, image similarity measurement is an important technique which can be applied in lots of applications, such as image retrieval, image matching, image recognition and classification, computer vision, etc. Many researchers work for this issue and numerous approaches have been proposed. For example, Coltuc et al. [24] studied the usefulness of the normalized compression distance (NCD for short) for image similarity detection. In their work, they considered correlation between NCD based feature vectors extracted for each image. Albanesi et al. [25] proposed a novel class of image similarity metrics based on a wavelet decomposition. They investigated the theoretical relationship between the novel class of metrics and the well-known structural similarity index. Abe et al. [26] studied similarity retrieval of trademark images represented by vector graphics. To improve the performance of the system, they introduced centroid distance into the feature extraction. Cicconet et al. [27] studied the problem of detecting duplication of scientific images. They introduced a data-driven solution based on a 3-branch Siamese Convolutional Neural Network which can serve to narrow down the pool of images. For multi-label image retrieval, Zhang et al. [28] proposed a novel deep hashing method named ISDH in which an instance-similarity definition was applied to quantify the pairwise similarity for images holding multiple class labels. Kato et al. [29] proposed a novel solutions for the problem of selecting image pairs that are more likely to match in Structure from Motion. They used Jaccard Similarity and bag-of-visual-words in addition to tf-idf to measure the similarity between images. Wang et al [30] designed a regularized distance metric framework which is named semantic discriminative metric learning (SDML for short). This framework combines geometric mean with normalized divergences and separates images from different classes simultaneously. Guha et al. [31] proposed a new approach called Sparse SNR (SSNR for short) to measuring the similarity between two images using sparse reconstruction. Their measurement does not need to use any prior knowledge about the data type or the application. KHAN et al. [32] proposed two halftoning methods to improve efficiency in generating structurally similar halftone images using Structure Similarity Index Measurement. Their Method I can improves efficiency as well as image quality and Method II can reaches a better image quality with fewer evaluations than pixel-swapping algorithm used in Method I.

Near-duplicate image detection is a another problem related to image similarity measurement. To solve the problem of near-duplicate image retrieval, Wang et al. [33] developed a novel spatial descriptor embedding method which encodes the relationship of the SIFT dominant orientation and the exact spatial position between local features and their context. Gadeski et al. [34] proposed an effective algorithm based on MapReduce framework to identify the near duplicates of images from large-scale image sets. Nian et al. [35] investigated this type of problem and presented an effective and efficient local-based representation method named Local-based Binary Representation to encode an image as a binary vector. Zlabinger et al. [36] developed a semi-automatic duplicate detection approach in which single-image-duplicates are detected between sub-images based on a connected component approach and duplicates between images are detected by using min-hashing method. Hsieh et al. [37] designed a novel framework that adopts multiple hash tables in a novel way for quick image matching and efficient duplicate image detection. Based on a hierarchical model, Li et al. [38] introduced an automatic NDIG mining approach by utilizing adaptive global feature clustering and local feature refinement to solve the problem of near duplicate image groups mining. Liu et al. [39] presented a variable-length signature to address the problem of near-duplicate image matching. They used the earth mover's distance to handle variable-length signatures. Yao et al. [40] developed a novel contextual descriptor which measures the contextual similarity of visual words to immediately discard the mismatches and reduce the count of candidate images. For large scale near-duplicate image retrieval Fedorov et al. [41] introduced a feature representation combining of three local descriptors, which is reproducible and highly discriminative. To improve the efficiency of near-duplicate image retrieval, Yıldız et al. [42] proposed a novel interest point selection method in which the distinctive subset is created with a ranking according to a density map.

Image Retrieval. Content-based image retrieval (CBIR for short) [43–46] is to retrieve images by analyzing visual contents, and therefore image representation [18,47] plays an important role in this task. In recent years, the task of CBIR has attracted more and more attentions in the multimedia [21,48,49] and computer vision community [19,20]. Many techniques have been proposed to support efficient multimedia query and image recognition. Scale Invariant Feature Transform (SIFT for short) [50,51] is a classical method to extract visual features, which transforms an image into a large collection of local feature vectors. SIFT includes four main step: (1) scale-space extrema detection; (2) keypoint localization; (3) orientation assignment; (4) Kkeypoint descriptor. It is widely applied in lots of

doi:10.20944/preprints201909.0232.v1

Notation	Definition
$\mathcal{D}_I$	A given database of images
$\mathcal{I}_i$	The i-th image
$\mathcal{W}_i$	A visual words set
$\lvert\mathcal{W}\rvert$	The number of visual words in $\mathcal{W}$
w_1^i	The i-th visual word in the visual words set $\mathcal{W}_i$
λ_k	The similarity of two visual words
$\mathcal{P}_k = (w_k^i, w_k^j)$	The similar visual word pair
$\otimes$	The operator to generates the set of SVWPs
$\hat{\lambda}$	The similarity threshold of predefined
Ξ_i	The set of visual words weight
$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$	The image similarity measurement
μ_i	The similarity of visual word
$\boldsymbol{\phi}$	The network parameters

Table 1. The summary of notations

researches and applications. For example, Ke et al. [52] proposed a novel image descriptor named PCA-SIFT which combines SIFT techniques and principal components analysis (PCA for short) method. Mortensen et al. [53] proposed a feature descriptor that augments SIFT with a global context vector. This approach adds curvilinear shape information from a much larger neighborhood to reduce mismatches. Liu et al. [54] proposes a novel image fusion method for multi-focus images with dense SIFT. This dense SIFT descriptor can not only be employed as the activity level measurement, but also be used to match the mis-registered pixels between multiple source images to improve the quality of the fused image. Su et al. [55] designed a horizontal or vertical mirror reflection invariant binary descriptor named MBR-SIFT to solve the problem of image matching. Nam et al. [56] introduced a SIFT features based blind watermarking algorithm to address the issue of copyright protection for DIBR 3D images. Charfi et al. [57] developed a bimodal hand identification system based on SIFT descriptors which are extracted from hand shape and palmprint modalities.

Bag-of-visual-words [19,58,59](BoVW for short) model is another popular technique for CBIR and image recognition, which was first used in textual classification. This model is a technique to transform images into sparse hierarchical vectors by using visual words, so that a large number of images can be manipulated. Santos et al. [60] presented the first ever method based on the signature-based bag of visual words (S-BoVW for short) paradigm that considers information of texture to generate textual signatures of image blocks for representing images. Karakasis et al. [61] presents an image retrieval framework that uses affine image moment invariants as descriptors of local image areas by BoVW representation. Wang et al. [62] presented an improved practical spatial weighting for BoV (PSW-BoV for short) to alleviate this effect while keep the efficiency.

3. Preliminaries

In this section, we propose the definition of region of visual interests (RoVI for short) at the first time, then present the notion of region of visual interests query (RoVIQ for short) and the similarity measurement. Besides, we review the techniques of image retrieval which is the base of our work. Table 1 summarizes the notations frequently used throughout this paper to facilitate the discussion.

3.1. Problem Definition

Definition 1 (Image object). *Let $\mathcal{D}_{\mathcal{I}}$ be an image dataset and $\mathcal{I}_i$ and $\mathcal{I}_j$ be two images, $\mathcal{I}_i, \mathcal{I}_j \in \mathcal{D}_{\mathcal{I}}$. We define the image object represented by bag-of-visual-word model as $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$, wherein $\mathcal{W}_i = \{w_1^i, w_2^i, ..., w_m^i\}$ and $\mathcal{W}_j = \{w_1^j, w_2^j, ..., w_n^j\}$ are the visual word set generated by feature extraction from $\mathcal{I}_i$ and $\mathcal{I}_j$, $\lvert\mathcal{W}_i\rvert = m$ and $\lvert\mathcal{W}_j\rvert = n$ are the number of visual words in these two sets respectively. In this study, we utilize image object as the representation model of images for the task of image similarity measurement.*

Definition 2 (Similarity of visual word). *Given two image objects $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$, wherein $\mathcal{W}_i = \{w_1^i, w_2^i, ..., w_m^i\}$ and $\mathcal{W}_j = \{w_1^j, w_2^j, ..., w_n^j\}$ are the visual words set. The similarity of two visual word $w_k^i \in \mathcal{W}_i$ and $w_k^j \in \mathcal{W}_j$ is represented by $\lambda_k = Sim_{\mathcal{W}}(w_k^i, w_k^j), \lambda_k \in [0,1]$, and if these visual words are identical, i.e., $\mathcal{W}_i = \mathcal{W}_j$, $\lambda_k = 1$.*

Definition 3 (Similar visual word pair). *Given two visual words $w_k^i \in \mathcal{W}_i$ and $w_k^j \in \mathcal{W}_j$ and the similarity of them is $\lambda_k = Sim_{\mathcal{W}}(w_k^i, w_k^j)$. Let $\bar{\lambda}$ is the similarity threshold predefined, if $\lambda_k > \bar{\lambda}$, this visual word pair is called as similar visual word pair (SVWP for short), represented as $\mathcal{P}_k = (w_k^i, w_k^j)$.*

Definition 4 (Similarity measurement of two image objects). *Given two image objects $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$. Let operation $\mathcal{W}_i \bigotimes \mathcal{W}_j = \{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_l\}$ generates the set of SVWPs which contain the visual words in $\mathcal{W}_i$ and $\mathcal{W}_j$, $l = |\mathcal{W}_i \bigotimes \mathcal{W}_j|$, and the similarity set of them are denoted as $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_l\}$, $\forall \lambda_i \in \Lambda, \lambda_i > \bar{\lambda}$. Let ξ_k^i and ξ_k^j be the weight of visual word w_k^i and w_k^j. For image objects $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$, the sets of their visual words weight are denoted as $\Xi_i = \{\xi_1^i, \xi_2^i, ..., \xi_l^i\}$ and $\Xi_j = \{\xi_1^j, \xi_2^j, ..., \xi_l^j\}$. The definitional equation of similarity between $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$ is shown as follows:*

$$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j)) = \mathcal{F}(m, n, l, \Lambda, \Xi_i, \Xi_j) \tag{1}$$

where m and n are the number of visual words of $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$ respectively. It is clearly that $Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ can meet the systematic similarity measurement criterion.

Theorem 1 (Monotonicity of similarity function). *The similarity measurement $Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ has the following five monotonicity conditions:*

- *$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ is a monotonic increasing function of weights of visual words in SVWPs, i.e., $\forall \xi_{w_x^i} \in \Xi_i$ and $\xi_{w_y^j} \in \Xi_j$, and $\forall \xi_{\hat{w}_x^i} \in \hat{\Xi}_i$ and $\xi_{\hat{w}_y^j} \in \hat{\Xi}_j$, if $\xi_{w_x^i} > \xi_{\hat{w}_x^i}$ and $\xi_{w_y^j} > \xi_{\hat{w}_y^j}$, $\mathcal{F}(m, n, l_1, \Lambda, \Xi_i, \Xi_j) > \mathcal{F}(m, n, l_2, \Lambda, \hat{\Xi}_i, \hat{\Xi}_j)$.*
- *$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ is a monotonic increasing function of the similarities of SVMPs $\Lambda = \{\lambda_1, \lambda_2, ...\lambda_l\}$, i.e., $\forall \lambda_x \in \Lambda$ and $\hat{\lambda_x} \in \hat{\Lambda}$, $\mathcal{F}(m, n, l_1, \Lambda, \Xi_i, \Xi_j) > \mathcal{F}(m, n, l_2, \hat{\Lambda}, \Xi_i, \Xi_j)$.*
- *$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ is a monotonic increasing function of number of SVWPs l, i.e., $\forall l_1, l_2 \in \mathbf{N}^+$, if $l_1 > l_2$, $\mathcal{F}(m, n, l_1, \Lambda, \Xi_i, \Xi_j) > \mathcal{F}(m, n, l_2, \Lambda, \Xi_i, \Xi_j)$.*
- *$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ is a monotonic decreasing function of weights of visual words which are not in SVWPs, i.e., .*
- *$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ is a monotonic decreasing function of the number of visual words which are not in SVWPs, i.e., if $r_1 = m + n - l_1$ and $r_2 = m + n - l_2$, $r_1 > r_2 \rightarrow l_1 < l_2$, $\mathcal{F}(m, n, l_1, \Lambda, \Xi_i, \Xi_j) < \mathcal{F}(m, n, l_2, \Lambda, \Xi_i, \Xi_j)$.*

According to the Definition 4 and theorem 1, the similarity measurement for two image objects is proposed, which is described in formal as follows.

Given two image objects $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$, $m = |\mathcal{W}_i|$ and $n = |\mathcal{W}_j|$. The sets of their visual words weight are $\Xi_i = \{\xi_1^i, \xi_2^i, ...\xi_l^i\}$ and $\Xi_j = \{\xi_1^j, \xi_2^j, ...\xi_l^j\}$. The SVMPs set of $\mathcal{I}_i(\mathcal{W}_i)$ and $\mathcal{I}_j(\mathcal{W}_j)$ is $\{\mathcal{P}_1, \mathcal{P}_2, ..., \mathcal{P}_l\}$, $l \leq min(m, n)$, and the similarities set of them is $\Lambda = \{\lambda_1, \lambda_2, ...\lambda_l\}$. The similarity measurement function $Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ is:

$$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j)) = \frac{\sum\limits_{k=1}^{l} \lambda_k \xi_k^i \xi_k^j}{\sqrt{\sum\limits_{k=1}^{m} \xi_k^i \sum\limits_{k=1}^{n} \xi_k^j} \sqrt{\sum\limits_{k=1}^{l} {\lambda_k}^2 \xi_k^i \xi_k^j + \sum\limits_{k=l+1}^{m} \xi_k^i \sum\limits_{k=l+1}^{n} \xi_k^j}} \tag{2}$$

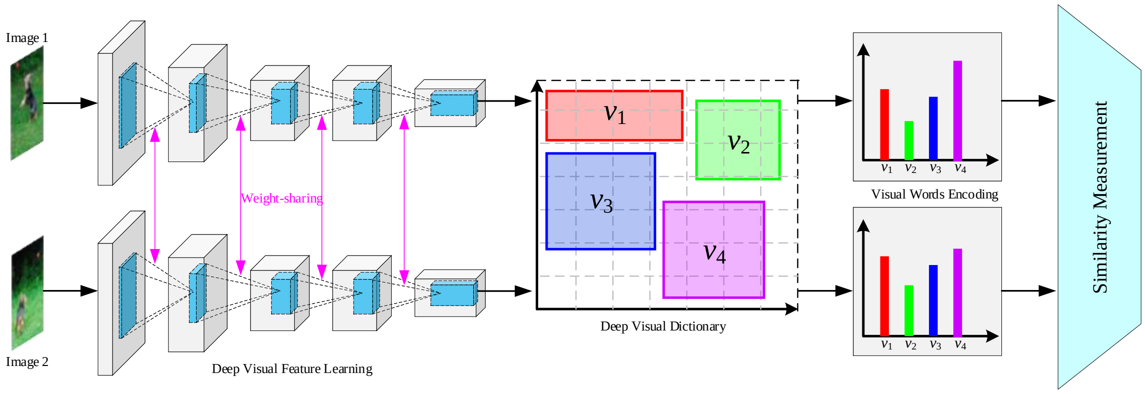


Figure 3. The framework of image similarity measurement. This framework employs a coupled CNNs network to learn the deep visual feature representations of the two input images. Across each layers of two CNNs, weight-sharing strategy is used to (1) learn the co-occurrence visual patterns, and (2) reduce the number of model parameters. Based on these visual representations, a deep visual dictionary is built by k-means method, which is used to encode the input images. After the generation of visual word representations of inputs, the proposed image similarity measurement is used to measure the visual similarity between the two input images.

Function 2 apparently meet the monotonicity described in Theorem 1. On the other hand, if these two image objects are identical, i.e., $\mathcal{I}_i(\mathcal{W}_i) = \mathcal{I}_j(\mathcal{W}_j)$, $\mathcal{W}_i = \mathcal{W}_j$, $m = n = l$, and $\xi_k^i = \xi_k^j$, then $Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j)) = 1$.

Theorem 2 (dissatisfying commutative law). *The similarity measurement* $Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j))$ *dissatisfy commutative law, i.e.,*

$$Sim_{\mathcal{I}}(\mathcal{I}_i(\mathcal{W}_i), \mathcal{I}_j(\mathcal{W}_j)) \neq Sim_{\mathcal{I}}(\mathcal{I}_j(\mathcal{W}_j), \mathcal{I}_i(\mathcal{W}_i))$$

In general, some visual words (e.g., noise words) in image objects have negative or reverse effects on the expression of the whole image. The SMI has a penalty effect on non-similar visual elements according to Theorem 1. this feature of the SMI has high accuracy for the similarity measurement of images.

4. Image Similarity Measurement Method

In this section, an efficient image similarity measurement framework via deep visual words is proposed, which is a combination of deep visual feature learning and Bag-of-Visual-Words technique. Besides, the similarity measurement algorithm and the optimization technique are introduced.

4.1. The framework of image similarity measurement

To effectively measure the similarity between two images via similar visual words, we carefully design a framework of image similarity measurement by combining the deep learning technique and BoVW model. Instead of the traditional visual words representation via SIFT descriptor and BoVW, we propose to use CNN to generate deep visual representations of images. Comparing to the traditional manner, this scheme can capture the rich high-level semantic concepts, which is more powerful for the image similarity measurement. Specifically, this framework uses a coupled CNNs network structure that recieve two input images, $\mathcal{I}_1$ and $\mathcal{I}_2$, and generate θ-dimensional deep visual representations, i.e., $(\zeta_1^{(1)}, \zeta_1^{(2)}, ..., \zeta_1^{(\theta)}) = Cnn_1(\mathcal{I}_1; \boldsymbol{\phi})$, $(\zeta_2^{(1)}, \zeta_2^{(2)}, ..., \zeta_2^{(\theta)}) = Cnn_2(\mathcal{I}_2; \boldsymbol{\phi})$, where $(\zeta_1^{(1)}, \zeta_1^{(2)}, ..., \zeta_1^{(\theta)})$ and $(\zeta_2^{(1)}, \zeta_2^{(2)}, ..., \zeta_2^{(\theta)})$ are the visual feature vector of $\mathcal{I}_1$ and $\mathcal{I}_2$, $\boldsymbol{\phi}$ is the network parameter. To learn co-occurrence visual patterns between the two inputs, a weight-sharing strategy is employed bewteen these two CNNs. On the other hand, weight-sharing can reduce the number of the network parameters significantly. For deep visual dictionary construction, k-means method is utilized to cluster these deep feature vectors into k groups, i.e., $k\text{-}Means(\{(\zeta^{(1)}, \zeta^{(2)}, ..., \zeta^{(\theta)})\}_n) = \{G\}_m$, where n is the number of input images, m is the number of groups. Accroding to the deep visual dictionary, the input images are encoded into visual words representations, i.e., $(\xi_1^i, \xi_2^i, ..., \xi_m^i) = TF\text{-}IDF(w_1^i, w_2^i, ..., w_m^i)$, $(\xi_1^j, \xi_2^j, ..., \xi_m^j) = TF\text{-}IDF(w_1^j, w_2^j, ..., w_m^j)$, where $TF\text{-}IDF$ is used to calculate the weight of each visual word, w is a weighted visual word. After the visual words representation generation, the paired visual words vectors are fed into the similarity measurement module to measure the visual similarity via similar visual words, which is discussed in Section 4.2 and 4.3.

In this work, we utilize pre-trained CNN model, AlexNet [63], to construct the coupled CNNs network. This network consists of five convolutional layers, three fully-connected layers and a 1000-way softmax layer. The 5-th convolutional representations 13 × 13 × 256 are used as the visual feature vectors for visual words generation. Besides, the input images are resized as 227 × 227 pixels.

4.2. The Measurement of Similar Visual Words

SMI is subject to the time complexity of the measurement of similar visual words. μ_i represents the similarity of a similar visual word as shown in the following formula:

$$\mu_i = \begin{cases} \arg\max_{b_j \in S_B} Sim_{\mathcal{I}}(a_i, b_j), & if > \mu_0 \\ 0, & if < \mu_0 \end{cases} \tag{3}$$

where $Sim_{\mathcal{I}}(a_i, b_j)$ represents the cosine of the angle between two vectors as the measurement of similarity. μ_0 is a judgment of the similarity threshold.

We give an intuitive way to measure similar visual words. The pseudo-code of the algorithm is shown in Algorithm 1. In this work, the double loop cosine method is called to be SMI Naive (SMIN for short).

Algorithm 1 SMIN Algorithm

```
Input $S_A$, $S_B$, $\mu_0$.
Output: $\mu$.
Initializing: $\mu \leftarrow \emptyset$;
Initializing: $S \leftarrow \emptyset$;
Initializing: $N_S \leftarrow \emptyset$;
Initializing: $maxsim \leftarrow 0$;
for each $\mathcal{W}_i \in S_A$ do
    for each $\mathcal{W}'_j \in S_B$ do
        if $cos(\mathcal{W}_i, \mathcal{W}'_j)$ then
            $maxsim \leftarrow cos(\mathcal{W}_i, \mathcal{W}'_j)$;
        end if
        if $maxsim \geq \mu_0$ then
            $S.Add(\mathcal{W}_i)$;
            $\mu.Add(maxsim)$;
        else
            $NS.Add(\mathcal{W}_i)$;
            $\mu.Add(0)$;
        end if
    end for
end for
return $\mu$;
```

4.3. The Optimization of Calculating Similar Visual Words

In the context of massive multimedia data, the multimedia retrieval system or image similarity measurement system requires an efficient sentence similarity measurement algorithm, the time complexity of the SMI focuses on the optimization of calculating similar visual words.

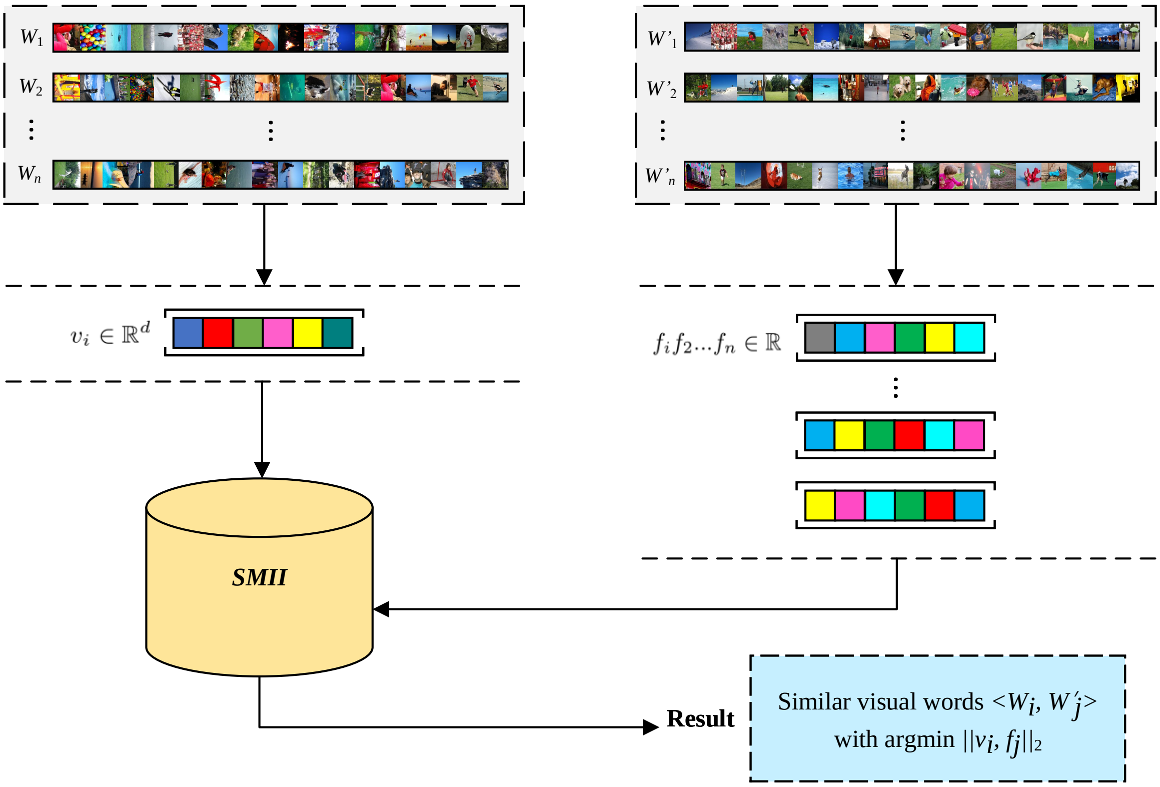

W_1
W_2
W_n
W'_1
W'_2
W'_n
$v_i \in \mathbb{R}^d$
$f_i f_2 \ldots f_n \in \mathbb{R}$
SMII
Result
Similar visual words <W_i, W'_j>
with argmin $||v_i, f_j||_2$

Figure 4. The processing of similarity measurement via SMII

SMI Temp Index. To reduce the double loop cos calculation to 1 cycle, a further approach is to construct an index γ_i of S_B for each vector a_i in S_A. According to experience, the dimension of the visual word vector is generally 200-300 dimensions to get better results.

For a vector a_i in S_A, we search for the vector b_j with the highest similarity in the temp index γ_i, so that the process requires only one similarity calculation. The n times calculations of similar visual words $< a_i, b_j >$ are reduced to vector searching, thereby reducing the execution time of *SMI*. However, there is a flaw that when every time a similar element of a sentence is calculated, a temp index needs to be built once, and the index cannot be reused. The temp index approach is called to be SMI Temp Index (SMII for short), as shown in Figure 4.

Index of Potential Similar Visual Words. In order to solve the problem that the index cannot be reused, we establish an index of potential similar visual words off-line in the process of word vector training. We could search for the index to perform the measurement of similar visual words without having to repeatedly create a temporary index. The main steps for index of potential similar visual words construction is shown as follows:

- Establishing an index for all the visual word vector set by trained word vector model.
- Traversing any vector $\mathfrak{v}$ to search the index to get a return set. In this set, the potential similar visual words are abstained with the similarity is greater than the threshold μ_0, in similarity descending order.
- The physic indexing structure of potential similar visual words could be implemented by a Huffman tree.

According to the hierarchical Softmax strategy in Word2Vec, an original Word2Vec Huffman tree constructed on the basis of the visual words frequency, and each node (except the root node) represents a visual word and its corresponding vector.

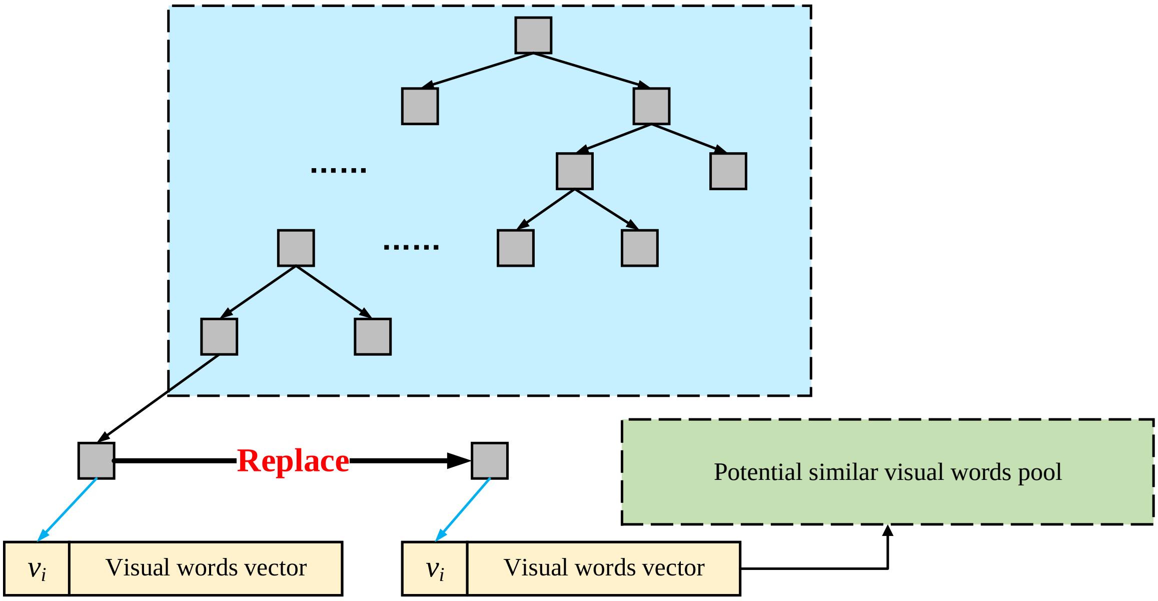


Figure 5. The index structure of potential similar visual words

We try to replace the vector with potential similar visual words. Thus each node of tree represents a visual word and its corresponding potential similar visual words. The index structure is illustrated by Figure 5:

We call the methods using global index of potential similar visual words as PSMI. Algorithm 2 illustrates the pseudo-code of PSMI.

Algorithm 2 PSMI Algorithm

```
Input: $S_A$, $S_B$, $\mu_0$
Output: $\mu$.
Initializing: $\mu \leftarrow \emptyset$;
Initializing: $S \leftarrow \emptyset$;
Initializing: $NS \leftarrow \emptyset$;
Initializing: $\mathcal{P} \leftarrow \emptyset$;
Initializing: $maxsim \leftarrow 0$;
for each $\mathcal{W}_i \in S_A$ do
    $\mathcal{P} \leftarrow HuffmanSearch(\mathcal{W}_i)$;
    for each $\mathcal{W}'_j \in S_B$ do
        for each $p_k \in \mathcal{P}$ do
            if $\mathcal{W}_j.equal(p_k.vector)$ then
                $S.Add(\mathcal{W}_i)$;
                $\mu.Add(p_k.sim)$;
                Break to loop $\mathcal{W}_i$;
            end if
        end for
    end for
    $NS.Add(\mathcal{W}_i)$;
    $\mu.Add(0)$;
end for
return $\mu$;
```

Algorithm 2 demonstrates the processing of the PSMI Algorithm. Firstly, for each visual words vector $\mathcal{W}_i \in S_A$, the algorithm executes the procedure $HuffmanSearch(\mathcal{W}_i)$ to get the node of the huffman tree which contains $\mathcal{W}_i$ and stored it in $\mathcal{P}$. Then, for each $\mathcal{W}'_j \in S_B$, the algorithm select each p_k from $\mathcal{P}$ and check if $\mathcal{W}_j$ is equal to $p_k.vector$ or not. if them are equal, the algorithm adds $\mathcal{W}_i$ into set S and adds $p_k.sim$ into μ. Then break to the outer loop. If $\mathcal{W}_i$ and $p_k.vector$ are not equal, then adds $\mathcal{W}_i$ into set NS and add 0 into μ.

Samples of Flickr

Samples of ImageNet

Figure 6. Some samples of Flickr and ImageNet dataset

4.4. Time complexity analysis of SMIN, SMII, PSMI

Suppose that the number of image pairs to be measured is Γ, the average number of visual word vectors of the visual word vector set S_A is $\bar{S_A}$, and the average number of visual word vectors of the visual word vector set S_B is $\bar{S_B}$, m represents the dimension of the vector.

For SMIN. Whether all elements which constitute similar visual words μ_i are calculated once by using formula, and the time consumption of calculation is determined by the number of vector dimension, the time complexity of SMIN is $O(\Gamma * \bar{S_A} * \bar{S_B} * COS)$, wherein COS is the time of cosine function $cos(.)$ between vector $\mathcal{W}_i$ and $\mathcal{W}_j'$, which equals to d. Thus the time complexity of SMIN is $O(\Gamma * \bar{S_A} * \bar{S_B} * m)$.

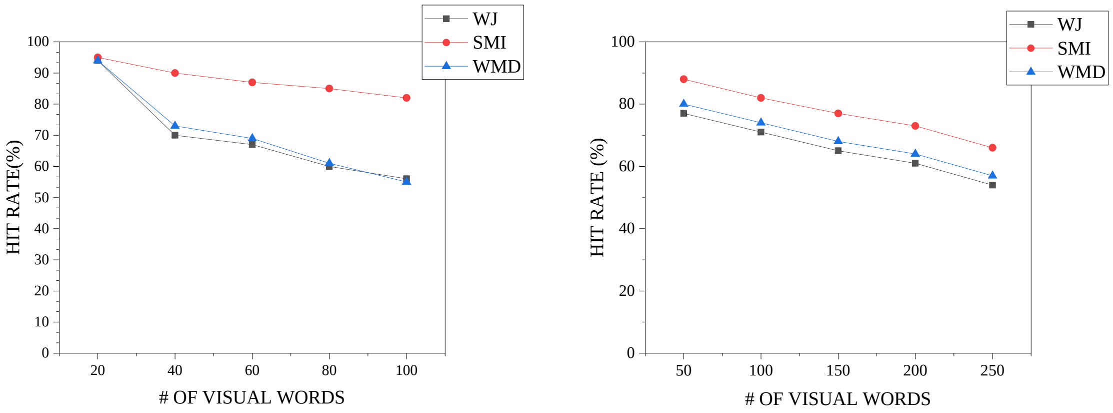
For SMII. To reduce the number of similar visual word calculations in method SMIN, the method of constructing an index is used, the index is equivalent to fuzzy search, and then the similar element is calculated to determine whether it constitutes a similar element, the time complexity of SMII is $O(\Gamma * \bar{S_A} * INDEX + \Gamma * \bar{S_B} * log(\bar{S_A}) * m)$, where $INDEX$ is the time to index each word vector, and $log(\bar{S_A})$ is the number of times to look up in the index. Since the $INDEX$ value is small, $\Gamma * \bar{S_A} * INDEX$ can be ignored. The time complexity of SMII approximately equals to $O(\Gamma * n * log(\bar{S_A}) * m)$.

For PSMI. PSMI constructs the Huffman tree offline, in which the potential similar elements of all word vectors are stored, and only the similar elements are calculated by searching. Thus, the time complexity of PSMI is $O(\Gamma * \bar{S_B} * log(|\mathcal{D}_I|))$, wherein $|\mathcal{D}_I|$ is the total number of the dictionary.

5. PERFORMANCE EVALUATION

In this section, we present results of a comprehensive performance study on real image datasets Flickr and ImageNet to evaluate the efficiency and scalability of the proposed techniques. Specifically, we evaluate the effectiveness of the following indexing techniques for region of visual interests search on road network.

- **WJ** WJ is the word2Vec technique proposed in https://github.com/jsksxs360/Word2Vec. In our experiments, we modify this technique for visual words.
- **WMD** WMD is the word2Vec technique, which is based on moving distance, is proposed in https://github.com/crtomirmajer/wmd4j.
- **SMIN** SMIN is the double loop cosine calculation technique proposed in Section 4.
- **SMII** SMII is the advanced technique of SMIN, which is proposed in Section 4.



(a) Evaluation on Flickr

(b) Evaluation on ImageNet

Figure 7. Evaluation on the number of visual words on Flickr and ImageNet

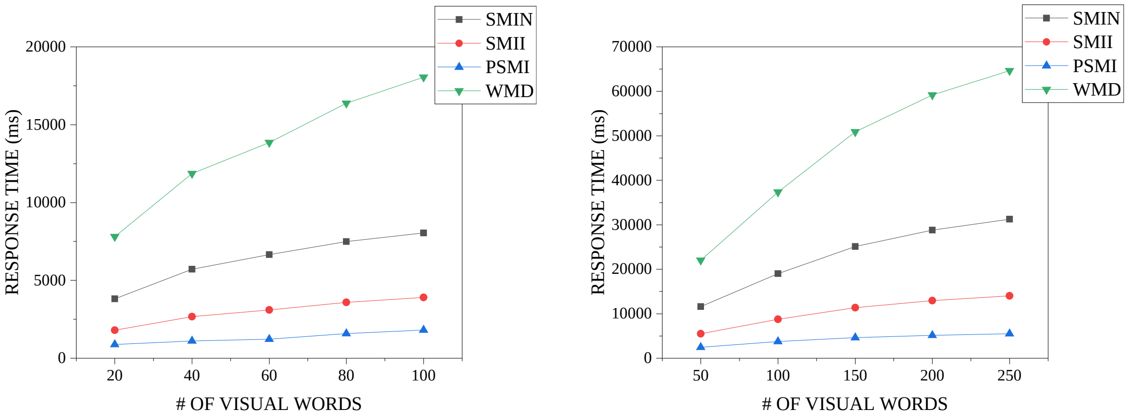
- **PSMI** PSMI is the potential similar visual words technique of SMII, which is also proposed in Section 4.

Datasets. Performance of various algorithms is evaluated on both real image datasets. We first evaluate these algorithms on **Flickr** is obtained by crawling millions image the photo-sharing site Flickr(http://www.flickr.com/). For the scalability and performance evaluation, we randomly sampled five sub datasets whose sizes vary from 200,000 to 1000,000 from the image dataset. Similarly, another image dataset **ImageNet**, which is widely used in image processing and computer vision, is used to evaluate the performance of these algorithms. Dataset **ImageNet** not only includes 14,197,122 images, but also contained 1.2 million images with SIFT features. We generate **ImageNet** dataset with varying size from 20K to 1M. Some samples of these two datasets are shown in Figure 6.

Workload. A workload for the region of visual interests query consists of 100 queries. The accuracy of these algorithm and the query response time is employed to evaluate the performance of the algorithms. The image dataset size grows from 0.2M to 1M; the number of the query visual words of dataset **Flickr** changes from 20 to 100; the number of the query visual words of dataset **ImageNet** varies from 50 to 250. The image dataset size, the number of the query visual words of dataset **Flickr**, and the number of the query visual words of dataset **ImageNet** set to 0.2M, 40, 100 respectively. Experiments are run on a PC with Intel Xeon 2.60GHz dual CPU and 16G memory running Ubuntu. All algorithms in the experiments are implemented in Java. Note that we only consider the algorithms WJ, SMI, WDM in accuracy comparison, because the SMIN, SMII, PSMI algorithms have the same error tolerance.

Evaluating hit rate on the number of visual words. We evaluate the hit rate on the number of query visual words on Flickr and ImageNet dataset shown in Figure 9. The experiment on Flickr is shown in Figure 9(a). It is clear that the hit rate of WJ, SMI and WMD decrease with the rising of the number of visual words. Specifically, the hit rate of our method, SMI, is the highest all the time. It descends slowly from around 90% to about 85%. On the other hand, the hit rate of WJ and WMD are very close. In the interval $[20, 40]$, they go down rapidly and after that the decrement of them become moderate. At 100, the hit rate of WJ is a litter higher than WMD, and both of them are much lower than SMI. In Figure 9(b), all of the decreasing trends are similar. Apparently, the hit rate of SMI is the highest, which goes down gradually with the increasing of the number of visual words. On ImageNet dataset, the hit rate of WMD is a litter higher than WJ all the time.

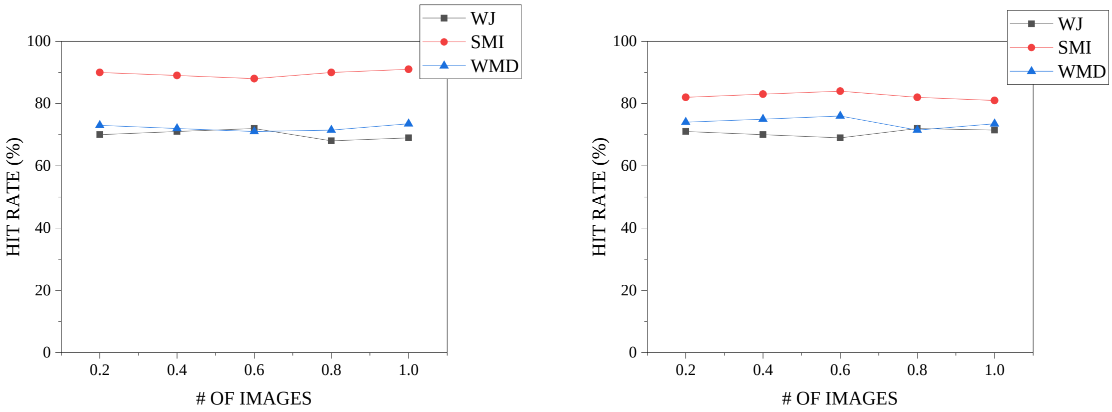
Evaluating response time on the number of visual words. We evaluate the response time on the number of visual words on Flickr and ImageNet dataset shown in Figure 10. In Figure 10(a), with the increment of number of visual words, the response time of PSMI has a slight growth, which is the lowest in these methods. The increasing trends of SMII is very moderate too, but it is slightly inferior



(a) Evaluation on Flickr

(b) Evaluation on ImageNet

Figure 8. Evaluation on the number of visual words on Flickr and ImageNet



(a) Evaluation on Flickr

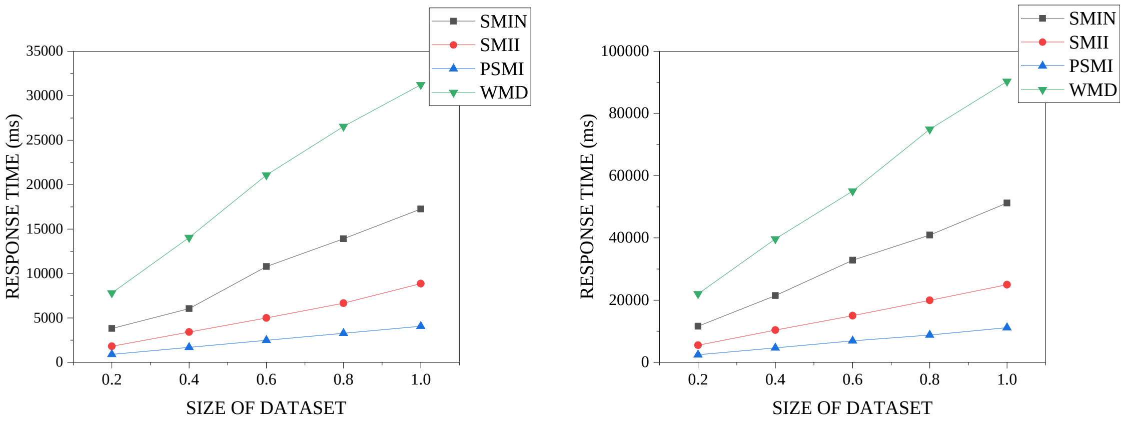
(b) Evaluation on ImageNet

Figure 9. Evaluation on the number of images on Flickr and ImageNet

to PSMI. Like PSMI and SMII, the performance of SMIN shows a moderate decrement with the rising of spatial similarity threshold. Although the response time of it is higher than the former two, it is much lower than WMD which has a fast growth in the interval of 20, 100. Figure 10(b) illustrates that the efficiency of PSMI is almost the same with the increment of number of visual words, which is the highest amount these four methods. Like the experiment on Flickr, the performance of both SMII and SMIN increase gradually and they are much better than WMD.

Evaluating hit rate on the number of images. We evaluate the hit rate on the number of images on Flickr and ImageNet dataset shown in Figure 9. Figure 9(a) demonstrates clearly that the hit rate of SMI is much higher than WJ and WMD. With the increasing of images number, it fluctuates slightly. the hit rate of WMD is almost unchanged with the increasing of number of images. On the other hand, the hit rate of WJ shows a moderate growth in the interval of 0.2, 0.6 and after that it drops and it is a litter lower than WMD. Clearly, the performance of SMI is the best. Figure 9(b) shows that the hit rate of SMI grows slightly in $[0.2, .06]$ and then go down weakly, which is higher than two others. Like the trend of SMI, the hit rate of WMD hit the maximum value at 0.6 and after that it decreases in the interval of $[0.6, 0.8]$. With this just the opposite is that the hit rate of WJ has a moderate decrement in $[0.2, 0.6]$ and it rises after 0.6.

Evaluating response time on the number of images. We evaluate response time on different size of query region on Flickr and ImageNet dataset shown in Figure 10. We can find from Figure 10(a)



(a) Evaluation on Flickr

(b) Evaluation on ImageNet

Figure 10. Evaluation on the number of images on Flickr and ImageNet

that the response time of PSMI and SMII increase slowly with the increasing of size of dataset. Both of them are much better than the others. The growth rate of SMIN is a litter higher than the two formers. The efficiency time of WMD is the worst. It grows rapidly and at 1.0 it is more than 30000ms. In Figure 10(b), we see that the growth of WMD is the fastest too. Like the situation on Flickr, the performance of WMD is the worst among them. By comparison, the upward trends of SMII and PSMI are much more moderate, and PSMI shows the best performance.

6. Conclusion

In this paper, we investigate the problem of image similarity measurement that is a significant issue in many applications. Firstly we proposed the definition of image objects and similarity measurement of two images and related notions. Then, an efficient image similarity measuremnt framework is proposed, which is a combination of a coupled CNNs network, BoVW model and similarity measurement via similar visual words. Based on Word2Vec, we develop the basic method of image similarity measurement, named SMIN. To improve the performance of similarity calculation, we improve this method and propose SMI Temp Index. To solve the problem of that the index cannot be reused, we develop a novel indexing technique called Index of Potential Similar Visual Words (PSMI). The experimental evaluation on real geo-multimedia dataset shows that our solution outperforms the state-of-the-art method.

Author Contributions: conceptualization, J.L. and L.Z.; methodology, L.Z.; software, X.Y. and L.S.; validation, L.Z., X.Y. and L.S.; formal analysis, L.Z.; investigation, L.Z. and X.Y; writing–original draft preparation, L.Z.; writing–review and editing, J.L.; visualization, L.S.; supervision, J.L.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China (61702560), project (2018JJ3691, 2016JC2011, 2016JC2018) of Science and Technology Plan of Hunan Province, the Science and Technology Plan of Hunan under Grant 2016TP1003, and the Research and Innovation Project of Central South University Graduate Students(2018zzts177,2018zzts588).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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.
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. Zhang, C., Lin, Y., Zhu, L., Liu, A., Zhang, Z., & Huang, F. (2019). CNN-VWII: An efficient approach for large-scale video retrieval by image queries. Pattern Recognition Letters, 123, 82-88.
4. 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.
5. 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.
6. 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.
7. Long, J., Zhu, L., Yang, Z., Zhang, C., & Yuan, X. (2018). Temporal Activity Path Based Character Correction in Heterogeneous Social Networks via Multimedia Sources. Advances in Multimedia, 2018.
8. 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.
9. 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.
10. Wu, L., Wang, Y., Gao, J., & Li, X. (2018). Where-and-when to look: Deep siamese attention networks for video-based person re-identification. IEEE Transactions on Multimedia, 21(6), 1412-1424.
11. 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.
12. 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.
13. 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.
14. Guo, S., & Xing, D. (2016). Sentence similarity calculation based on word vector and its application research [D]. Modern Electronics Technique.
15. Li F., Hou J., Zeng R., & Ling C. (2017). Research on Multi-Feature Sentence Similarity Computing Method with Word Embedding [J]. Journal of Frontiers of Computer Science and Technology.
16. 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.
17. Zhang, C., Wu, L., & Wang, Y. (2019). Crossing generative adversarial networks for cross-view person re-identification. Neurocomputing, 340, 259-269.
18. 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.
19. 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.
20. 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, 29(10), 4833-4843.
21. Wu, L., Wang, Y., Li, X., & Gao, J. (2018). Deep attention-based spatially recursive networks for fine-grained visual recognition. IEEE transactions on cybernetics, 49(5), 1791-1802.
22. Zhang, C., Cheng, K., Zhu, L., Chen, R., Zhang, Z., & Huang, F. (2018). Efficient continuous top-k geo-image search on road network. Multimedia Tools and Applications, 1-30.
23. 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.
24. Coltuc, D., Datcu, M., & Coltuc, D. (2018). On the use of normalized compression distances for image similarity detection. Entropy, 20(2), 99.
25. Albanesi, M. G., Amadeo, R., Bertoluzza, S., & Maggi, G. (2018). A new class of wavelet-based metrics for image similarity assessment. Journal of Mathematical Imaging and Vision, 60(1), 109-127.
26. Abe, K., Morita, H., & Hayashi, T. (2018, February). Similarity Retrieval of Trademark Images by Vector Graphics Based on Shape Characteristics of Components. In Proceedings of the 2018 10th International Conference on Computer and Automation Engineering (pp. 82-86). ACM.

27. Cicconet, M., Elliott, H., Richmond, D. L., Wainstock, D., & Walsh, M. (2018). Image Forensics: Detecting duplication of scientific images with manipulation-invariant image similarity. arXiv preprint arXiv:1802.06515.
28. Zhang, Z., Zou, Q., Wang, Q., Lin, Y., & Li, Q. (2018). Instance similarity deep hashing for multi-label image retrieval. arXiv preprint arXiv:1803.02987.
29. Kato, T., Shimizu, I., & Pajdla, T. (2017). Selecting image pairs for SfM by introducing Jaccard Similarity. IPSJ Transactions on Computer Vision and Applications, 9(1), 12.
30. Wang, H., Feng, L., Zhang, J., & Liu, Y. (2016). Semantic discriminative metric learning for image similarity measurement. IEEE Transactions on Multimedia, 18(8), 1579-1589.
31. Guha, T., Ward, R. K., & Aboulnasr, T. (2013, May). Image similarity measurement from sparse reconstruction errors. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 1937-1941). IEEE.
32. Khan, A., Aguirre, H., & Tanaka, K. (2012). Improving the Efficiency in Halftone Image Generation Based on Structure Similarity Index Measurement. IEICE TRANSACTIONS on Information and Systems, 95(10), 2495-2504.
33. Wang, Y., & Zhou, Z. (2018). Spatial descriptor embedding for near-duplicate image retrieval. International Journal of Embedded Systems, 10(3), 241-247.
34. Zhao, W., Luo, H., Peng, J., & Fan, J. (2017). MapReduce-based clustering for near-duplicate image identification. Multimedia Tools and Applications, 76(22), 23291-23307.
35. Nian, F., Li, T., Wu, X., Gao, Q., & Li, F. (2016). Efficient near-duplicate image detection with a local-based binary representation. Multimedia Tools and Applications, 75(5), 2435-2452.
36. Zlabinger, M., & Hanbury, A. (2017, April). Finding duplicate images in biology papers. In Proceedings of the Symposium on Applied Computing (pp. 957-959). ACM.
37. Hsieh, S. L., Chen, C. C., & Chen, C. R. (2015). A novel approach to detecting duplicate images using multiple hash tables. Multimedia Tools and Applications, 74(13), 4947-4964.
38. Li, J., Qian, X., Li, Q., Zhao, Y., Wang, L., & Tang, Y. Y. (2015). Mining near duplicate image groups. Multimedia Tools and Applications, 74(2), 655-669.
39. Liu, L., Lu, Y., & Suen, C. Y. (2015). Variable-length signature for near-duplicate image matching. IEEE Transactions on Image Processing, 24(4), 1282-1296.
40. Yao, J., Yang, B., & Zhu, Q. (2014). Near-duplicate image retrieval based on contextual descriptor. IEEE signal processing letters, 22(9), 1404-1408.
41. Fedorov, S., & Kacher, O. (2016). Large scale near-duplicate image retrieval using Triples of Adjacent Ranked Features (TARF) with embedded geometric information. arXiv preprint arXiv:1603.06093.
42. Yıldız, B., & Demirci, M. F. (2016, March). Distinctive interest point selection for efficient near-duplicate image retrieval. In 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) (pp. 49-52). IEEE.
43. Jing, Y., & Baluja, S. (2008). Visualrank: Applying pagerank to large-scale image search. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11), 1877-1890.
44. Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2(1), 1-19.
45. 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.
46. 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.
47. 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.
48. 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.
49. Zhang, C., Lin, Y., Zhu, L., Zhang, Z., Tang, Y., & Huang, F. (2018). Efficient region of visual interests search for geo-multimedia data. Multimedia Tools and Applications, 1-25.

50. Lowe, D. G. (1999, September). Object recognition from local scale-invariant features. In iccv (Vol. 99, No. 2, pp. 1150-1157).
51. Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.
52. Ke, Y., & Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. CVPR (2), 4, 506-513.
53. Mortensen, E. N., Deng, H., & Shapiro, L. (2005, June). A SIFT descriptor with global context. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 184-190). IEEE.
54. Liu, Y., Liu, S., & Wang, Z. (2015). Multi-focus image fusion with dense SIFT. Information Fusion, 23, 139-155.
55. Su, M., Ma, Y., Zhang, X., Wang, Y., & Zhang, Y. (2017). MBR-SIFT: A mirror reflected invariant feature descriptor using a binary representation for image matching. PloS one, 12(5), e0178090.
56. Nam, S. H., Kim, W. H., Mun, S. M., Hou, J. U., Choi, S., & Lee, H. K. (2018). A SIFT features based blind watermarking for DIBR 3D images. Multimedia Tools and Applications, 77(7), 7811-7850.
57. Charfi, N., Trichili, H., Alimi, A. M., & Solaiman, B. (2017). Bimodal biometric system for hand shape and palmprint recognition based on SIFT sparse representation. Multimedia Tools and Applications, 76(20), 20457-20482.
58. Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. In null (p. 1470). IEEE.
59. Wu, L., Wang, Y., & Shao, L. (2018). Cycle-consistent deep generative hashing for cross-modal retrieval. IEEE Transactions on Image Processing, 28(4), 1602-1612.
60. Dos Santos, J. M., De Moura, E. S., Da Silva, A. S., & da Silva Torres, R. (2017). Color and texture applied to a signature-based bag of visual words method for image retrieval. Multimedia Tools and Applications, 76(15), 16855-16872.
61. Karakasis, E. G., Amanatiadis, A., Gasteratos, A., & Chatzichristofis, S. A. (2015). Image moment invariants as local features for content based image retrieval using the bag-of-visual-words model. Pattern Recognition Letters, 55, 22-27.
62. Wang, F., Wang, H., Li, H., & Zhang, S. (2013, January). Large scale image retrieval with practical spatial weighting for bag-of-visual-words. In International Conference on Multimedia Modeling (pp. 513-523). Springer, Berlin, Heidelberg.
63. Krizhevsky, A. , Sutskever, I. , & Hinton, G. . (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS (Vol.25). Curran Associates Inc.