Using Word Mover's Distance with Spatial Constraints for Measuring Similarity Between Mongolian Word Images

Hongxi Wei(✉), Hui Zhang, Guanglai Gao, and Xiangdong Su

School of Computer Science, Inner Mongolia University, Hohhot 010021, China
cswhx@imu.edu.cn

Abstract. In the framework of bag-of-visual-words, visual words are independent each other, which results in discarding spatial relations and lacking semantic information of visual words. To capture semantic information of visual words, a deep learning procedure similar to word embedding technique is used for mapping visual words to embedding vectors in a semantic space. And then, word mover's distance (WMD) is utilized to measure similarity between two word images, which calculates the minimum traveling distance from the visual embeddings of one word image to another one. Moreover, word images are partitioned into several sub-regions with equal sizes along rows and columns in advance. After that, WMDs can be computed from the corresponding sub-regions of the two word images, separately. Thus, the similarity between the two word images is the sum of these WMDs. Experimental results show that the proposed method outperforms various baseline and state-of-the-art methods, including spatial pyramid matching, latent Dirichlet allocation, average visual word embeddings and the original word mover's distance.

Keywords: Visual word embeddings · Word mover's distance · Spatial information · Keyword spotting · Query-by-example

1 Introduction

How to access the content from a large number of scanned historical document images is still a challenging task. Because of aging, the historical document images are often degradation and poor quality. Therefore, robust optical character recognition (OCR) tools are not available yet. When OCR is infeasible, keyword spotting technology is an alternative approach. In the keyword spotting technology, all scanned historical document images are generally segmented into individual word images to form a word image collection. As for a given query keyword, relevant word images can be detected in the collection of word images by image matching [1].

In the traditional keyword spotting, profile-based features were widely used to represent word images [2] and dynamic time warping (DTW) algorithm was utilized to accomplish image matching [3]. Though the DTW algorithm works well, it is so time-consuming that cannot be suited for real-time image matching for a large collection of word images. Hence, this study focuses on how to represent word images so as to realize real-time image matching.

D. Liu et al. (Eds.): ICONIP 2017, Part IV, LNCS 10637, pp. 192–201, 2017.
https://doi.org/10.1007/978-3-319-70093-9_20

In recent years, Bag-of-Visual-Words (BoVW) has been attracted much more attention and shown advantages in keyword spotting on historical documents [4, 5]. In the BoVW framework, word images are represented as visual histograms with a fixed-length. In this way, cosine similarity (or Euclidean distance) between word images can be calculated on their histograms. At the retrieval stage, when a query keyword is provided, the corresponding cosine similarities can be calculated for a collection of word images. By this way, a ranking list of word images can be formed in descending order of the cosine similarities. Thus, the BoVW-based representation approach is competent for the task of keyword spotting on a large number of word images. However, local descriptors (i.e. visual words) within one word image are independent each other in the BoVW-based representation, which results in not only discarding spatial relations between visual words but also lacking semantic information of visual words.

In this paper, an approach has been proposed to capture semantic information of visual words. To be specific, a deep learning procedure similar to word embedding is used for mapping visual words to embedding vectors in a semantic space. Consequently, the semantic relatedness between visual words can be measured by calculating Euclidean distance or cosine similarity on their embedding vectors. In order to distinguish from the original word embeddings proposed by Mikolov et al. [6], the embedding vectors in this study are called *visual word embeddings*. Kusner et al. [7] recently proposed word mover's distance (WMD), a distance function between two documents, which calculates the minimum traveling distance from the word embeddings of one document to another one. In our study, the WMD is used for measuring the similarity between two word images. Through this way, the semantic information of visual words is integrated into image matching.

Additionally, all word images are partitioned into a certain quantity of sub-regions with equal sizes along rows and columns in advance. In the image matching phase, only the corresponding sub-regions of the two word images are matched each other. Thus, WMDs can be computed from the corresponding sub-regions of the two word images, separately. Finally, the similarity of the two word images is the sum of these WMDs. By this means, such the spatial relations can be added to the procedure of image matching. Hence, the above-mentioned drawbacks of the BoVW-based representation can be overcome using the proposed WMD with spatial constraints.

The rest of the paper is organized as follows. The related work is presented in Sect. 2. The proposed method is described detailedly in Sect. 3. Experimental results are shown in Sect. 4. Section 5 provides the conclusions and future work.

2 Related Work

In the keyword spotting technology, several manners for providing query keywords have been proposed in the literature, which can be divided into query-by-example (QBE) and query-by-string (QBS) approaches [8]. In the QBS approach, query keywords are provided by textual strings [9, 10]. But, the QBS approach needs to learn a model to map from textual strings to images on a certain number of annotated word images. When there is no such annotated word images, the QBE approach can be

competent. The QBE approach [11, 12] requires that an example image of a query keyword is provided for being retrieved. In this study, we concentrate on the QBE based approach for realizing keyword spotting on historical Mongolian document images.

In our previous work, visual language model (VLM) [13] was proposed for representing the corresponding word images segmented from a collection of historical Mongolian documents. Therein, each word image was represented as a probability distribution of visual words and query likelihood model was used to calculate similarity between two word images. Although the VLM (e.g. bigram visual language model) can provide the spatial orders between the neighboring visual words, there is still lacking semantic information of visual words. Therefore, a latent Dirichlet allocation (LDA) based word image representation approach presented in another our previous work [14]. In the LDA-based representation, topics were treated as probability distributions over visual words. Each word image was viewed as a probabilistic mixture over these topics. Thus, the LDA-based representation can provide the semantic information of visual words. However, the semantic information in the LDA-based representation is latent, which cannot be used for measuring the semantic relatedness between visual words directly. Consequently, the semantic information needs to be obtained in more obvious form.

In the last few years, word embedding techniques have been shown significant improvements in various natural language processing (NLP) tasks, such as word analogy [6], information retrieval [15], and so forth. Word2vec [6] and GloVe [16] are examples of successful implementations of word embeddings that respectively use neural networks and matrix factorization to learn embedding vectors. Because GloVe incorporates co-occurrence statistics of words that frequently appear together within the documents. Pennington et al. [16] have proved that GloVe outperforms Word2vec on word analogy, word similarity and named entity recognition tasks. Therefore, GloVe is utilized to generate embedding vectors for visual words in this paper. In this manner, visual words will be mapped as vectors in a semantic space. The generated embedding vectors are named *visual word embeddings*. As far as we know, this is the first time to learn and generate embedding vectors on visual words.

After that, a common approach for representing a word image is to take a centroid of its visual word embeddings. And then, an inner product or cosine between the centroids can be calculated for measuring similarity [17]. However, taking a simple centroid is not a good approximation for representing a word image. A more reasonable approach is to calculate similarity between visual words from one word image to another one. Consistent with this, Kusner et al. [7] proposed a word mover's distance that can calculate similarity between two documents on their embedding vectors. This study is partly motivated by the word mover's distance to measure similarity between two word images.

To integrate the spatial relations of visual words into image matching, spatial pyramid matching (SPM) has been proposed by Lazebnik et al. [18]. The SPM method partitions an image into several sub-regions with equal sizes and computes visual histograms in each sub-region. Our study is also inspired by the SPM method. All word images are partitioned into a number of sub-regions with equal sizes along rows and columns. At the stage of image matching, only the corresponding sub-regions of the two word images are matched each other. In this paper, similarity between two word images is measured by using the WMD with spatial constraints.

3 Proposed Method

In our study, the handling objects are word images. Hence, each scanned image in a collection of historical Mongolian documents should be segmented into individual word images in advance. And the QBE approach is used in the retrieval phase. The details of the proposed method are presented in the following subsections.

3.1 Obtaining Visual Words

Given a collection of word images, SIFT descriptors are extracted from each word image. And then, *k-means clustering algorithm* is applied on these SIFT descriptors so as to generate a certain number of clusters. Thus, the center of each cluster is taken as a visual word. By this way, a codebook can be formed. Figure 1 shows the procedure for constructing a codebook.

After that, each SIFT descriptor will be assigned the label of the closet center (i.e. visual word) according to the codebook. Thus, the corresponding visual words can be obtained from the collection of word images.

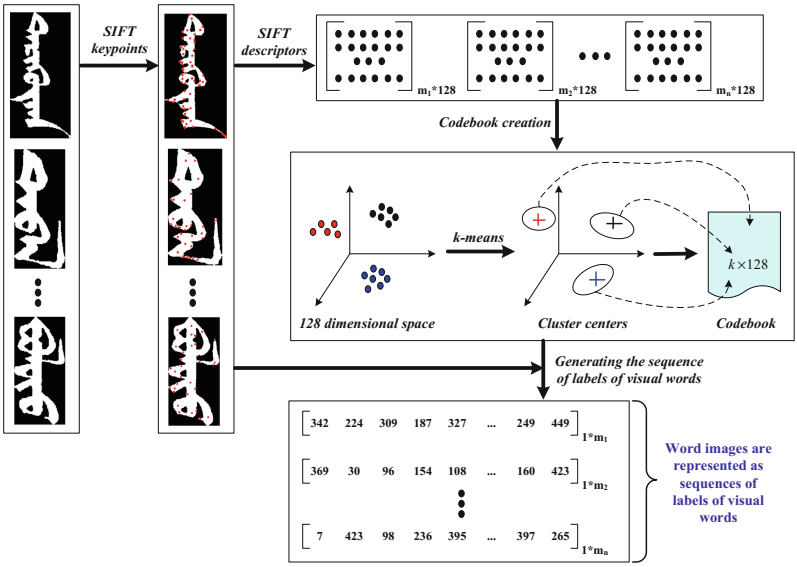

Fig. 1. The procedure for constructing a codebook.

3.2 Generating Visual Word Embeddings

After obtaining visual words, one word image can be represented as a sequence of labels of visual words along the writing direction (see Fig. 1). On a collection of word images, a training corpus of visual words can be collected by concatenating the corresponding sequences of labels of visual words one after another. And then, a GloVe tool (http://nlp.stanford.edu/projects/glove/) is utilized to generate embedding vectors of visual words on the training corpus.

In our study, the parameters of GloVe are set to as follows. The size of embedding vector and context window are set to 200 and 15, separately. And the number of iterations is set to 15. After generating visual word embeddings, the semantic relatedness between visual words can be measured by calculating Euclidean distance (or cosine similarity) on their embedding vectors.

3.3 Word Mover's Distance

Our work is based on the original word mover's distance (WMD) between text documents proposed by Kusner et al. in [7]. The average time complexity of the original WMD is $O(n^3 \log n)$, where n denotes the number of vocabularies on a collection of documents. For documents with many unique words, solving the WMD optimal transport problem may become prohibitive. So, Kusner et al. also introduced relaxed and much faster WMD versions.

In our case, the first relaxation is to sum the distances of the visual word embeddings $\vec{w}$ in a query keyword q to the closest visual word embeddings $\vec{w}'$ of a word image d. Thus, the WMD from the query keyword q to the word image d (denoted by RWMD_{Q2D}) can be defined as follows.

$$\mathrm{RWMD}_{Q2D}(q \to d) = \sum\nolimits_{w \in q} \frac{count(w)}{\sum_{t \in q} count(t)} . \min_{w' \in d} distance\left(\vec{w}, \vec{w}'\right) \tag{1}$$

where w and w' are the visual words occurred in q and d, separately. $\vec{w}$ and $\vec{w}'$ are the corresponding visual embedding vectors of w and w'. $count(w)$ means the occurrence frequency of the visual word w in q. And $\sum_{t \in q} count(t)$ means the total number of visual words in q. $distance\left(\vec{w}, \vec{w}'\right)$ denotes the Euclidean distance between two visual embedding vectors $\vec{w}$ and $\vec{w}'$, and its formulation is as follows.

$$distance\left(\vec{w}, \vec{w}'\right) = \sqrt{\sum\nolimits_{i=1}^{K} (w_i - w_i')^2} \tag{2}$$

where w_i and w_i' denote the i^{th} elements in the visual embedding vectors $\vec{w}$ and $\vec{w}'$. K indicates the dimension of the visual word embeddings. In our study, K equals to 200.

Similarly, the second relaxed form is to sum the distances of the visual word embeddings $\vec{w}'$ of d to the closest visual word embeddings $\vec{w}$ of q. The corresponding WMD from the word image d to the query keyword q (denoted by RWMD_{D2Q}) can be defined as the following formula.

$$\mathrm{RWMD}_{D2Q}(d \to q) = \sum\nolimits_{w' \in d} \frac{count(w')}{\sum_{t' \in d} count(t')} . \min_{w \in q} distance\left(\vec{w}', \vec{w}\right) \tag{3}$$

In (1) and (3), the time complexity for getting the optimal solution is $O(n^2)$, which is faster than the original WMD. Kusner et al. found the maximum of RWMD_{Q2D} and

RWMD_{D2Q} to be the best relaxation of the original WMD. Therefore, the final WMD between two word images q and d can be calculated by the following equation.

$$\mathrm{WMD}(q,d) = \max\left\{\mathrm{RWMD}_{Q2D}(q \to d),\ \mathrm{RWMD}_{D2Q}(d \to q)\right\} \tag{4}$$

In this way, when a query keyword and a collection of word images are given, a ranking list of word images can be formed according to (4).

3.4 Integrating Spatial Information

In order to integrate spatial information into word images representation, all word images are partitioned into a quantity of sub-regions along rows and columns. All sub-regions within one word image have the equal sizes. Figure 2 depicts an example for partitioning a word image into three sub-regions along rows.

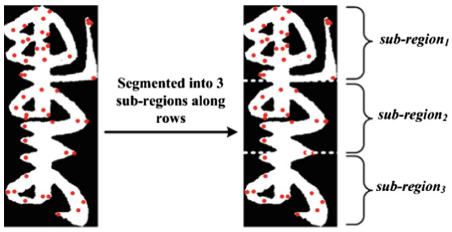

Fig. 2. An example for partitioning a word image.

At the image matching phase, the corresponding sub-regions between two word images are matched, respectively. Thus, the RWMD_{Q2D} and RWMD_{D2Q} between two word images can be rewritten as follows.

$$\mathrm{RWMD}_{Q2D}(q \to d) = \sum_{j=1}^{N} RWMD_{Q2D}\left(q_j \to d_j\right) \tag{5}$$

$$\mathrm{RWMD}_{D2Q}(d \to q) = \sum_{j=1}^{N} RWMD_{D2Q}\left(d_j \to q_j\right) \tag{6}$$

where q_j and d_j denote the j^{th} sub-regions of the two word images, severally. N indicates the number of sub-regions. Particularly wish to point out, the difference between the adopted spatial information in this study and the SPM is regardless of partition levels.

4 Experimental Results

4.1 Dataset and Baselines

To evaluate the performance, a collection of Mongolian historical documents has been collected, which consists of **100** scanned Mongolian Kanjur images with **24,827** words.

Each page has been annotated manually to form the ground truth data. Twenty meaningful words are selected and taken as query keywords. The dataset and the query keywords are the same as in [13, 14]. Evaluation metric is *mean average precision* (MAP).

For constructing a codebook, SIFT descriptors are extracted from the **24,827** word images and the total number of the SIFT descriptors is **2,283,512**. After that, the k-means clustering algorithm has been performed on those descriptors. Therein, we vary the number of clusters from **500** to **10,000** with **500** as an interval. In the following subsections, the appropriate number of clusters will be determined.

In this section, *spatial pyramid matching*, *average visual word embeddings*, *visual language model* and *latent Dirichlet allocation* are taken as baselines for comparison. The details of the baseline methods are as follows.

Spatial pyramid matching (SPM): The standard SPM method [18] is utilized to accomplish the aim of image matching between a given query keyword and each word image in a collection.

Average visual word embeddings (AVWE): After generating visual embedding vectors, a word image (denoted by W) can be represented as a centroid (denoted by W_{cent}) of the embedding vectors of its visual words using the following equation [17]:

$$W_{cent} = \frac{1}{|W|} \sum_{j=1}^{|W|} v_j \tag{7}$$

where v_j is the embedding vector of the corresponding visual word and $|W|$ is the number of visual words within the word image W. Under the circumstance, Euclidean distance can be calculated and used for measuring similarity between word images.

Visual language model (VLM): Each word image can be represented as probability distribution of visual words. Query likelihood model is utilized to rank word images. In our previous work [13], the best performance of VLM can attain to **31.75%**.

Latent Dirichlet allocation (LDA): A LDA-based topic model is adopted to obtain the semantic relations between visual words. In our previous work [14], the best performance is **43.78%**. At present, the LDA-based representation method is the state-of-the-art for keyword spotting on the same dataset.

4.2 Performance of SPM and AVWE

For Mongolian language, its writing direction is from top to bottom. So, horizontal partitions are more important than vertical partitions. In the SPM method, we have tested nine types of one-level partitions and five types of two-level partitions. Their MAPs are shown in Figs. 3 and 4, respectively.

From Figs. 3 and 4, we can see that one-level partitions and two-level partitions both obtain the best performance when the number of clusters is **500**. In various one-level partitions, the best performance is **37.71%** and the manner of the partition is **9 * 2**. Correspondingly, the best performance of the two-level partitions is **38.38%** when the

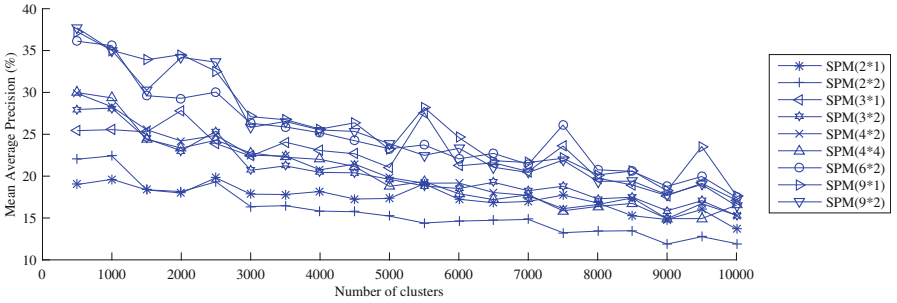

Fig. 3. The performance of SPM with one-level partitions.

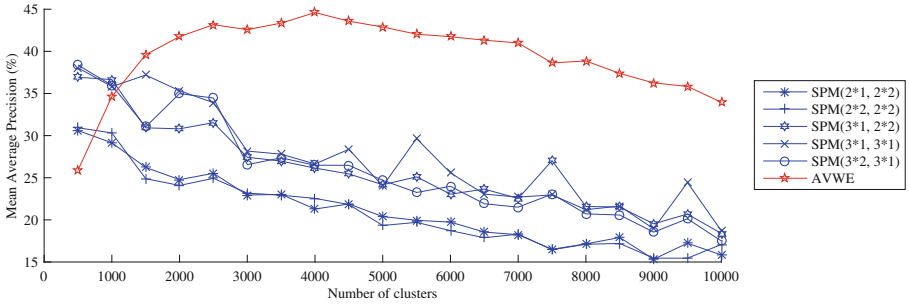

Fig. 4. The performance of SPM with two-level partitions.

first level and the second level are **3 * 2** and **3 * 1**, severally. Indeed, the horizontal partitions are more crucial than the vertical partitions for Mongolian word images.

Additionally, the performance of AVWE has been tested. In Fig. 4, the best performance of AVWE is **44.61%** when the number of clusters is **4,000**. Therefore, the AVWE is superior to the SPM. The MAP is improved from 38.38% to 44.61%. It indicates that the semantic information of visual words is more important than the spatial information in our case. Meanwhile, the AVWE is superior to the LDA-based method as well. So, the proposed visual embeddings can capture much more semantic information than the LDA-based representation method.

4.3 Performance of the Relaxed WMD Without Spatial Constraints

For comparison, we have also tested the performance of the relaxed WMD without spatial constraints. According to (1), (3) and (4), word images can be ranked for a given query keyword. In Fig. 5, the best performance is **38.16%** when the number of clusters is **7500**. Although the relaxed WMD is superior to the VLM (31.75%) and the one-level partition based SPM (37.71%), it is inferior to the other baseline methods including the tow-level partition based SPM (38.38%), the LDA-based representation method (43.78%) and the AVWE (44.61%).

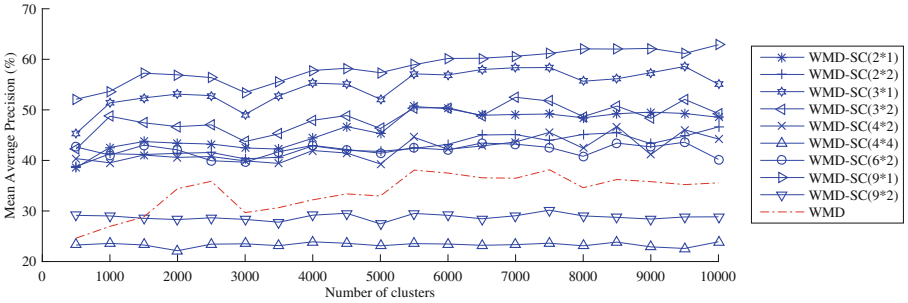

Fig. 5. The performance of the relaxed WMD and the proposed WMD-SC.

4.4 Performance of the Proposed WMD with Spatial Constraints

Here, we tested the performance of the proposed relaxation WMD with spatial constraints (denoted by **WMD-SC**). There are nine partition styles for word images, which are the same as in Fig. 3. The corresponding results of the proposed **WMD-SC** are shown in Fig. 5. As for the proposed **WMD-SC**, the various partition styles are consistently superior to the WMD without spatial constraints except for 4 * 4 and 9 * 2. Therefore, the spatial information plays an important part in our study.

In Fig. 5, the best performance of the **WMD-SC** is **62.88%** when the number of clusters is **10,000** and word images are partitioned into **9** sub-regions along rows only. Therefore, the performance of the proposed **WMD-SC** is increased by **44%** (from 43.78% to 62.88%) against to the state-of-the-art method (i.e. LDA-based representation method) on the same dataset.

5 Conclusion and Future Work

In this paper, a novel method has been proposed for measuring similarity between Mongolian word images. On the one hand, the spatial information is obtained by partitioning word images into sub-regions along rows and columns. Only the corresponding sub-regions between two word images are matched. On the other hand, embedding vectors of visual words are generated by utilizing a deep learning tool. After that, the relaxed word mover's distance is used for calculating similarity on the corresponding sub-regions between two word images. Therefore, the proposed method can combine the spatial information of visual words with the semantic relatedness so as to attend the aim of measuring similarity between word images. And the performance of the proposed method outperforms various baseline methods and the state-of-the-art method.

In our future work, the corresponding semantic relatedness between the visual embeddings will be utilized to attain the aim of query expansion. The proposed method will be validated on the other datasets of historical documents.

Acknowledgement. This paper is supported by the National Natural Science Foundation of China under Grant 61463038.

References

1. Rath, T.M., Manmatha, R.: Word spotting for historical manuscripts. Int. J. Doc. Anal. Recogn. **9**(2), 139–152 (2007)
2. Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Proceedings of ICDAR 2003, pp. 218–222. IEEE Press, New York (2003)
3. Rath, T.M., Manmatha, R.: Word image matching using dynamic time warping. In: Proceedings of CVPR 2003, pp. 521–527. IEEE Press, New York (2003)
4. Shekhar, R., Jawahar, C.V.: Word image retrieval using bag of visual words. In: Proceedings of DAS 2012, pp. 297–301. IEEE Press, New York (2012)
5. Aldavert, D., Rusinol, M., Toledo, R., Llados, J.: A study of bag-of-visual-words representations for handwritten keyword spotting. Int. J. Doc. Anal. Recogn. **18**(3), 223–234 (2015)
6. Mikolov, T., Sutskever, I., Chen, K., Coorado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS 2013, pp. 3111–3119. MIT Press, Massachusetts (2013)
7. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. Proc. Mach. Learn. Res. **37**, 957–966 (2015)
8. Fornes, A., Frinken, V., Fischer, A., Almazan, J., Jackson, G., Bunke, H.: A keyword spotting approach using blurred shape model-based descriptors. In: Proceedings of HIP 2011, pp. 83–89. ACM Press, New York (2011)
9. Aldavert, D., Rusinol, M., Toledo, R., Llados, J.: Integrating visual and textual cues for query-by-string word spotting. In: Proceedings of ICDAR 2013, pp. 511–515. IEEE Press, New York (2013)
10. Rothacker, L., Fink, G.A.: Segmentation-free query-by-string word spotting with bag-of-features HMMs. In: Proceedings of ICDAR 2015, pp. 661–665. IEEE Press, New York (2015)
11. Wei, H.X., Gao, G.L., Su, X.D.: A multiple instances approach to improving keyword spotting on historical Mongolian document images. In: Proceedings of ICDAR 2015, pp. 121–125. IEEE Press, New York (2015)
12. Wei, H.X., Zhang, H., Gao, G.L.: Representing word image using visual word embeddings and RNN for keyword spotting on historical document images. In: Proceedings of ICME 2017, pp. 1374–1379. IEEE Press, New York (2017)
13. Wei, H.X., Gao, G.L.: Visual language model for keyword spotting on historical Mongolian document images. In: Proceedings of CCDC 2017, pp. 1765–1770. IEEE Press, New York (2017)
14. Wei, H., Gao, G., Su, X.: LDA-based word image representation for keyword spotting on historical Mongolian documents. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) ICONIP 2016. LNCS, vol. 9950, pp. 432–441. Springer, Cham (2016). doi:10.1007/978-3-319-46681-1_52
15. Zamani, H., Croft, W.B.: Embeddings-based query language models. In: Proceedings of ICTIR 2016, pp. 147–156. ACM Press, New York (2016)
16. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of EMNLP 2014, pp. 1532–1543. ACL Press, Stroudsburg (2014)
17. Nalisnick, E., Mitra, B., Craswell, N., Caruana, R.: Improving document ranking with dual word embeddings. In: Proceedings of WWW 2016, pp. 83–84. ACM Press, New York (2016)
18. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR 2006, pp. 2169–2178. IEEE Press, New York (2006)