Research Article

On Building a Universal and Compact Visual Vocabulary

Jian Hou,¹ Wei-Xue Liu,¹ Xu E,¹ and Hamid Reza Karimi²

¹ School of Information Science and Technology, Bohai University, Jinzhou 121013, China

² Department of Engineering, Faculty of Engineering and Science, University of Agder, 4898 Grimstad, Norway

Correspondence should be addressed to Jian Hou; dr.houjian@gmail.com

Received 28 August 2013; Accepted 1 October 2013

Academic Editor: Zhiguang Feng

Copyright © 2013 Jian Hou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Bag-of-visual-words has been shown to be a powerful image representation and attained great success in many computer vision and pattern recognition applications. Usually, for a given dataset, researchers choose to build a specific visual vocabulary from the dataset, and the problem of deriving a universal visual vocabulary is rarely addressed. Based on previous work on the classification performance with respect to visual vocabulary sizes, we arrive at a hypothesis that a universal visual vocabulary can be obtained by taking-into account the similarity extent of keypoints represented by one visual word. We then propose to use a similarity threshold-based clustering method to calculate the optimal vocabulary size, where the universal similarity threshold can be obtained empirically. With the optimal vocabulary size, the optimal visual vocabularies of limited sizes from three datasets are shown to be exchangeable and therefore universal. This result indicates that a universal and compact visual vocabulary can be built from a not too small dataset. Our work narrows the gap between bag-of-visual-words and bag-of-words, where a relatively fixed vocabulary can be used with different text datasets.

1. Introduction

Bag-of-visual-words is a powerful and widely used image representation in computer vision and pattern recognition applications. In this approach, salient image regions (keypoints) are detected, described, and then clustered into groups. By treating the centroid of each group as a visual word, we obtain a visual vocabulary composed of all visual words. With this vocabulary, an image can be represented as a histogram of the visual words, namely, a bag-of-visual-words [1]. Since one visual word represents one type of image patterns, a bag-of-visual-words can be regarded as the distribution of various image patterns in an image. The basic bag-of-visual-words representation captures this distribution in the whole image and ignores the spatial relationships among keypoints, and this is shown to weaken the discriminative power of this representation [2, 3]. In order to make use of the spatial information, in [4] the authors proposed to use a spatial pyramidal partition of an image and concentrate the histogram in each partition into one final description. Some other approaches to encode spatial information include [5, 6] considering that some visual words in a visual vocabulary

may be more informative than the others in a specific domain, [3, 7, 8] propose to weight visual words accordingly, and [9, 10] present methods to reduce vocabulary size for better efficiency. Furthermore, it was proposed in [7] to build a vocabulary tree by hierarchical k -means clustering to scale to large vocabularies and large datasets.

While the potential of bag-of-visual-words has been explored in various aspects as we briefly reviewed above, the construction of visual vocabularies almost always follows a fixed procedure; that is, randomly selecting some images from the given datasets, extracting keypoints from these images, and building the specific vocabulary by clustering in the keypoints, where the vocabulary size is selected empirically and may range from hundreds to tens of thousands [1, 4, 11, 12]. This practice is quite different from that of bag-of-words in text domain, where a universal and limited vocabulary can be used for different datasets. Noticing that bag-of-words is the counterpart and precoder of bag-of-visual-words, we are interested to find out if it is possible to eliminate this difference and build a universal and compact visual vocabulary. By universal we mean that on different datasets, the visual vocabulary performs comparable to their

specifically trained vocabularies, and thereby removing the necessity to build a specific visual vocabulary for each dataset, and by *compact* we mean that the vocabulary is not so large, in case users feel it is not worthwhile to use a large vocabulary. An earlier version of some works in this paper appeared in [13].

In the literature, the most related works to ours are the ones in [14, 15] which address the problem of deriving a universal visual vocabulary. Specifically, it was empirically found in [14, 15] that the visual vocabularies trained from one dataset can be used on some other datasets without apparently harming the performance, only if the dataset is large enough. In these two papers, the vocabulary sizes are still user-defined and this implies that an inappropriate vocabulary size may lead to a universal vocabulary that performs moderately, which obviously is not what we really expect. On the contrary, in our work, the universal vocabulary is naturally derived from our work on optimal vocabulary size. Furthermore, the universal vocabulary obtained with our approach is optimal and compact, in that it can be used on different datasets to obtain the (near-)best performance, and the vocabulary size is only several thousands. These two important properties are not possessed by the universal vocabularies achieved in [14, 15].

It should be noted that in [16, 17], the term *universal* is used with different meanings. In [16], a universal vocabulary refers to a large vocabulary which serves as the basis of generating small and optimal vocabularies. Whereas in [17], a universal vocabulary is built from images from all categories, in contrast to the class-specific ones built from one category. On the contrary, in this paper, a vocabulary is universal in that it can be used with different datasets to achieve the (near-)best performance.

The remainder of this paper is organized as follows. In Section 2, we present details on using a similarity threshold-based clustering method and an optimal similarity threshold to determine the optimal vocabulary size for a given dataset. Then, in Section 3, we test if the optimal vocabulary trained from one dataset can be used to produce the (near-)best performance on other datasets, and thus decide if it can be used as a universal visual vocabulary. Section 4 discusses the experimental results in Sections 2 and 3 and their implications. Finally, Section 5 concludes the paper.

2. Optimal Visual Vocabulary

2.1. Hypothesis. This work originates from the observations from experiments on vocabulary sizes in [3], where it is found that with the increase of vocabulary size, the classification performance of the visual vocabulary rises dramatically until a peak is reached, and after that, the performance levels off or drops mildly. Experiments in [18] also showed that when the vocabulary size reaches a certain value, a larger size does not improve the performance further. These observations imply that the increase in vocabulary size does not definitely improve the classification performance. Instead, there exists an optimal vocabulary size for a given dataset, and the optimal size is smaller than the number of keypoints. This,

in turn, indicates that if a set of keypoints are similar enough to each other, they should be represented by one single visual word, not only for efficiency reasons, but also for the best performance. With a descriptor of limited dimension, for example, SIFT, this means that all the possible keypoints can be mapped to a limited number of visual words, which can then be used as a universal visual vocabulary. Here, by optimal we mean that the vocabulary performs the (near-)best and a larger size does not pay off.

We have speculated that there should exist a universal visual vocabulary of limited size. However, it is still not clear how to derive such a visual vocabulary. Noticing that a universal visual vocabulary is the optimal one for any dataset, our solution to universal visual vocabulary derivation is to build an optimal vocabulary from a large and representative dataset, and then test if it is also the optimal ones for other datasets. For a given dataset, the optimal vocabulary is decided by the optimal vocabulary size. We already know that there exists an optimal vocabulary size for a given dataset, but we do not know how to determine the optimal vocabulary size. Empirically selecting such a size is not a good option, because the optimal vocabulary size varies from dataset to dataset, and there are so many sizes to test for each dataset. In fact, we propose to use a similarity threshold-based clustering method and an optimal similarity threshold to determine the optimal vocabulary size for a given dataset. After the optimal vocabularies for some datasets are obtained by k -means clustering, we test if they are also the optimal ones for other datasets to decide if they can be treated as universal visual vocabularies.

Determining the optimal vocabulary size is the key step in building the optimal vocabulary from a given dataset. While [3, 18] show that there exists an optimal vocabulary size smaller than the number of training descriptors for a given dataset, they do not provide the reason behind the observation and provide a solution to find the optimal size. Instead, they just test many sizes and select the best performing one. In this paper, we intend to analyze the reasons behind the observations and then derive a method to determine the optimal visual vocabulary size.

2.2. Similarity Based Clustering Method. With the popular detectors, for example, DoG used in SIFT [19], usually at least hundreds of keypoints can be detected from one image. Therefore, with only hundreds of training images, we can obtain at least tens of thousands of training keypoints. In this case, using a vocabulary size of several hundreds implies that a large number of keypoints are clustered into one single group and are represented by one single visual word. This often means a large intracluster dissimilarity and leads to little discriminating power of the vocabulary. With the increase of vocabulary size, one visual word tends to represent only similar keypoints and the discriminating power of the vocabulary tends to increase. In this way, a larger vocabulary should perform better until the largest vocabulary size, that is, the number of keypoints, is reached. However, the experiments in [3, 18] show that when the vocabulary size is large enough (i.e., the optimal size observed), a larger size

does not result in performance gain. This further indicates that if a set of keypoints are similar enough, they should be clustered into one single group and represented by one single visual word, instead of many different visual words, respectively. This is a very important conclusion and it serves as the basis of deriving a universal visual vocabulary which is optimal and of limited size. Therefore, in the following, we explain it in a little more detail.

Firstly, at the optimal size, the keypoints represented by one visual word are so similar that they actually describe the same image pattern and their difference is too small to be taken into account. Secondly, the difference in keypoints represented by the same visual word may be caused by noise and thus should be ignored. This viewpoint is also supported by the work on local descriptors matching in [20] and visual words weighting scheme in [3]. In [20], we show that in straightforward local keypoints matching, the number of matched keypoint types is a better image similarity measure than that of the number of matched keypoints. Here, by type, we mean that one type of keypoints are very similar to each other and thus should be regarded as of one type. In [3], experiments on binary weighting and term-frequency weighting indicate that when the vocabulary size is large enough, the count of keypoints in one visual word provides no more useful information than the presence or absence of the visual word. If we regard the keypoints of enough similarity as of one type, all these works lead to the same conclusion that it is better to treat keypoints by types than individually. In other words, in bag-of-visual-words representation, the optimal vocabulary size should be equal to the number of training keypoints types, and not the number of training keypoints. Now what is left for us to do is to define when a set of keypoints can be regarded as of one type and calculate the number of keypoint types among keypoints.

Since the keypoints of one type are similar to each other, we use a similarity threshold to define the notion of keypoint type. Specifically, we regard a set of keypoints as of one type if their similarities with their mean are all above a threshold th_s . In order to calculate the number of keypoint types among training descriptors, we need to do clustering among keypoints where each cluster corresponds to one keypoint type. Besides, we want the number of clusters to be minimized so that all keypoints of one type are really grouped into one cluster. Since the number of clusters is not known beforehand, the k -means-like clustering methods cannot be used here. Based on our previous work [13, 21], we propose to use the following similarity threshold based clustering method.

- (1) Label all training keypoints as unclustered.
- (2) Label the first unclustered keypoint as the centroid of one cluster.
- (3) Compare each unclustered keypoint with the current center, and add it into the current cluster if the similarity is bigger than th_s .
- (4) Return to Step 2 until all keypoints are clustered.
- (5) Calculate the new centroid of each cluster, and use the count of keypoints in the cluster as the weight.
- (6) Sort the centroids by weight in decreasing order.

(7) Compare all keypoints with each centroid in order and add to the corresponding cluster if the similarity is bigger than th_s .

(8) If there are keypoints left unclustered, repeat Step 2 to 3 to cluster them into new clusters.

(9) Repeat Steps 5 to 8 for a certain time.

The above presented is a simple clustering procedure to serve our purpose of obtaining a clustering with minimal clusters and each cluster corresponds to one keypoint type. Although we do not have theoretical evidence that this clustering procedure is guaranteed to converge and minimize the number of clusters, all our experiments show that the number of clusters obtained decreases gradually and tends to be stable after 5 iterations. Recalling that the performance of bag-of-visual-words is not sensitive to small changes in vocabulary size, we stick to this procedure in this paper and use the results after 10 iterations in all experiments.

Different from the traditional clustering methods which strengthen high intracluster similarity and low intercluster similarity, our clustering method requires all clusters to have an identical distribution range determined by the similarity threshold. In other words, our clustering method actually seeks to obtain an even and disjoint partition of the keypoint feature space by a set of hyperspheres whose radii are determined by the similarity threshold. Since the features in each hypersphere are represented by visual words, the visual vocabulary obtained in this way covers the whole feature space and is guaranteed to be of the minimum size. This visual vocabulary can therefore be regarded as a universal one and used with any datasets. Here, we see that the similarity threshold based clustering method is not a clustering method in the strict sense, and it is just an approach to find the similarity extent of one visual word in the universal visual vocabulary.

2.3. Optimal Similarity Threshold. In order to determine the optimal vocabulary size, we still need an appropriate selection of the similarity threshold th_s . By definition, th_s decides if a set of keypoints should be treated as of one type. Theoretically deriving such a threshold, if feasible, is out of our scope in this paper as it may involve physiological and psychological issues. Instead, we choose to empirically select the optimal similarity threshold th_s . In the first step, we test several candidates of th_s on different datasets and check if the best performing candidates coincide with each other. If a unique best performing th_s is identified, the corresponding vocabulary size is then compared with other candidate sizes to see if it is the best performing one. It should be pointed out that in our experiments, the similarity threshold based clustering is only used to determine the vocabulary size, and the vector quantization with all vocabulary sizes is done with k -means clustering method. In this way, we make sure that the performance difference of different vocabulary sizes is not due to different clustering methods.

The experimental setup is as follows. DoG and SIFT [19] is used to detect and describe keypoints, and cosine similarity is selected as the keypoints similarity measure.

We do SVM classification on three diverse datasets: Caltech-101 [22], Scene-15 [4, 23, 24], and Event-8 [25]. With the well known Caltech-101, we randomly select 30 images per class as training and up to 15 images in the remaining as testing. Different from some of the literature using only 101 classes, here we test with all 102 classes. The Scene-15 dataset is composed of images of 15 classes with 200 to 400 images in each class. Following the setup in [4], we use 100 randomly selected images per class in training and all the others in testing. The Event-8 dataset contains images of 8 sports categories and each category has 130 to 250 images. This dataset is challenging for classification, not just in classifying events from static images, but that cluttered and diverse background, and various poses, sizes, and views of foreground objects are involved. We follow the setup in [25], that is, randomly selecting 70 images per class as training and another 60 images as testing.

We build the visual vocabularies of each dataset based on a set of randomly selected images from the dataset. In order to avoid any additional influence on the classification performance, the bag-of-visual-words histograms are built in the whole image, that is, at spatial pyramid level 0. In classification, we use term-frequency weighting scheme to build linear kernels. The multiclass SVM is trained in a one-versus-all setup and the regulation parameter C is fixed to be 1000. For all three datasets, we test with 10 training-testing splits and report the mean results. The performance measure adopted is the mean recognition rate per class. Note that in all our experiments, we use visual words without spatial information or special kernel; therefore, we do not expect to obtain superior classification performance comparable to the state-of-the-art. What is really important here is the trend of recognition rates with respect to the vocabulary sizes.

In the first step, we compare the performance of 4 candidate similarity thresholds 0.7, 0.75, 0.8, and 0.85. We do not adopt larger or smaller candidates as they produce extremely large or small vocabulary sizes that are obviously far from the optimal ones in our experiments. The four sizes calculated with similarity based clustering are 544, 2323, 12593, and 88328 for Caltech-101; 455, 1790, 9208, and 59539 for Event-8; and 560, 2378, 13124, and 92735 for Scene-15. The classification rates are reported in Figure 1, where we use similarity thresholds instead of the specific vocabulary size to show the trend more evidently. From this figure, we find that the vocabulary sizes from $th_c = 0.75$ and $th_c = 0.8$ perform the best.

We then compare the vocabulary sizes from these two optimal similarity thresholds with other sizes 100, 1000, 10000, 50000, and 100000 to check if they still perform the best. The results are shown in Table 1. As the optimal sizes corresponding to similarity thresholds 0.75 and 0.8 are different for different datasets, in the leftmost column we use $th_c = 0.75$ and $th_c = 0.8$ to represent their respective sizes.

It is evident from Table 1 that the vocabulary sizes from $th_c = 0.75$ and $th_c = 0.8$ perform the best or near-best among all sizes with three datasets. This confirms that $th_c = 0.75$ and $th_c = 0.8$ can be used to produce the optimal vocabulary size. Taking into account the small performance difference and large size difference between thresholds 0.8 and 0.75, we

FIGURE 1: Recognition rates of different similarity thresholds. With all three datasets, the similarity thresholds 0.75 and 0.8 produce the best or near-best performance.

TABLE 1: Classification rates of different vocabulary sizes on three datasets. The sizes corresponding to similarity thresholds 0.75 and 0.8 produce the best or near-best results in all cases.

	caltech-101	Event-8	Scene-15
100	36.6 ± 1.0	70.7 ± 2.5	65.0 ± 0.4
1000	48.5 ± 1.1	79.3 ± 2.1	73.1 ± 0.7
10000	48.3 ± 1.1	80.0 ± 2.2	73.9 ± 0.5
50000	47.1 ± 0.9	78.5 ± 2.1	72.1 ± 0.5
100000	46.5 ± 1.1	78.1 ± 2.0	71.3 ± 0.7
$th_c = 0.75$	49.5 ± 0.9	79.5 ± 2.4	73.9 ± 0.6
$th_c = 0.8$	49.0 ± 1.0	80.4 ± 2.5	73.8 ± 0.5

recommend to select 0.75 as the optimal similarity threshold in practical applications. Although Table 1 also indicates that the exact optimal similarity threshold should lie between these two values, we do not bother to seek it and are satisfied with these two candidates. One main consideration is that the performance of a visual vocabulary is not very sensitive to its size, only if the size is not too small. Our experiments indicate that for a common dataset of about the size of Caltech-101, 1000, or 2000, it might be a suitable vocabulary size. Adopting a larger size usually does not pay off.

In this, section we present a similarity threshold based clustering method to determine the optimal vocabulary size for a given dataset. Although the similarity threshold is determined empirically, this method cannot be regarded as just transforming the problem of determining k in k -means to the one of determining th_c . For different datasets the optimal vocabulary sizes are usually different, as shown in Table 1. This implies that one cannot select the optimal vocabulary size from one dataset and apply this size to other datasets.

Whereas with our similarity threshold based clustering method, the different optimal vocabulary sizes from different datasets correspond to a relatively fixed threshold th_s . This means that we can use this threshold to compute the optimal vocabulary size for a newly given dataset. More importantly, the results of this section support our hypothesis that the whole feature space can be mapped to a limited number of visual words, as one visual word is able to represent a set of keypoints that are similar enough. As shown in Section 3, this hypothesis is the basis of obtaining a universal and compact visual vocabulary.

3. Universal Visual Vocabulary

With a given dataset, researchers usually choose to build a specific visual vocabulary, instead of using an existing universal vocabulary. One possible explanation is that researchers tend to believe that the optimal visual vocabulary is data dependent and must be built specifically. However, the existence of an optimal visual vocabulary size smaller than the number of keypoints indicates that if some keypoints are similar enough, they should be grouped into one cluster and represented by one single visual word, but not by different visual words separately. This further implies that all keypoints can be mapped to a limited number of visual words. With the visual vocabulary obtained this way, all image patterns with semantic meanings can be represented with enough precision. This amounts to say that the images from any datasets can be represented accurately by this vocabulary; that is, this vocabulary is universal across datasets.

Recall that the optimal similarity threshold sets a criterion for keypoints to be mapped to one single visual word. Theoretically it is possible to enumerate all the possible visual words with the optimal similarity threshold th_s . However, it is not clear if all these image patterns represented by the vocabulary occur frequently in images or have semantic meanings. In other words, by enumerating all possible image pattern types, we may obtain a visual vocabulary that is complete but of a very large size. Nevertheless, many of these image patterns may rarely appear in real images. This causes unnecessary computation load and reduces the necessity of obtaining a universal vocabulary. Therefore, in this paper, we resort to empirical methods.

In Section 2, we have computed the optimal visual vocabularies for three datasets, which we refer to as *voc-caltech*, *voc-event*, and *voc-scene*, respectively. Here, we interchange the roles of datasets and visual vocabularies to check if different visual vocabularies produce a large performance difference on the same datasets. Take *voc-caltech* for example, we use it on *Event-8* and *Scene-15* and see if it performs comparably to *voc-event* and *voc-scene*, respectively. The comparison is shown in Figure 2.

Contrary to the traditional viewpoint that a good visual vocabulary is data dependent, we found from the comparison in Figure 2 that with each dataset, the visual vocabularies built with the three datasets perform rather similarly. This seems to imply that the visual vocabularies built from different datasets have a rather larger portion of visual words in

common. In order to validate this viewpoint, we calculate the pairwise similarity between three vocabularies. Specifically, for each visual word in one vocabulary, we compute its cosine similarity with its closest counterpart in the other vocabulary. For all the 6 cases, that is, *Caltech-Event*, *Event-Caltech*, *Caltech-Scene*, *Scene-Caltech*, *Event-Scene*, and *Scene-Event*, almost all visual words have $a > 0.9$ similarity with their counterparts in other vocabularies, and over 60% of the visual words have $a > 0.95$ similarity. These results confirm that the three vocabularies are very similar to each other. This is interesting as the descriptors from three diverse datasets are mapped to almost identical vocabularies with our optimal vocabulary sizes. This observation confirms our belief that there does exist a universal visual vocabulary and the difference in appearances of images is only caused by the different distribution of image patterns represented by these visual words in the vocabulary. In other words, the experiments indicate that obtaining a universal visual vocabulary is not only theoretically sound, but practically feasible.

In [14], the authors conclude through experiments that with a given vocabulary size large enough, the visual vocabularies built from different datasets are exchangeable without harming the classification performance evidently. Therefore, a large vocabulary needs to be computed only once and can be used as a universal one. However, it is not clear if the universal visual vocabulary obtained this way is optimal or suboptimal. A vocabulary that is universal but performs ordinarily is not what we need definitely. In this paper, we arrive at much more powerful conclusions. When we say an optimal vocabulary is universal, our meaning is threefold. Firstly, the vocabulary can be used on other datasets to obtain comparable performance as their specific vocabulary. Secondly, our vocabulary is optimal in that it can produce the best or near-best performance on any datasets. Thirdly, our optimal vocabularies are of limited size (1000 to several thousands). This not only means efficiency in classification, but implies that a very large vocabulary may not be necessary at all. To sum up, in this paper, we provide an approach to produce a vocabulary that is universal, optimal, and compact.

Although currently we only test the approach on three datasets, we also note that all the three datasets contain objects of diverse types and large variations, and are thus rather representative. In the next step, we will extend the experiments to more and larger datasets, for example, *Caltech-256* [26], *Oxford flowers* [27], *NUS-WIDE* [28], and *Graz* [29], and so forth, in order to finally produce a universal visual vocabulary, which can be used in a large number of datasets for the best or near-best performance.

4. Discussion

Our work in this paper is motivated from the difference between bag-of-words and bag-of-visual-words and also the experiments in [3]. Since bag-of-words is the counterpart and precursor of bag-of-visual-words, we expect them to share some common properties. However, we observe that in text domain, a universal vocabulary of relatively fixed size can

FIGURE 2: Recognition rates with vocabularies trained from different datasets. x-axis represents different testing datasets, and the different bars indicate vocabularies trained from different datasets. It is clear that different vocabularies perform similarly on the same datasets.

be used on different datasets, whereas in image domain the visual vocabularies are usually built specifically for given datasets and the vocabulary sizes are user-defined or determined empirically. On the other hand, the experiments in [3, 18] indicate that for an image dataset, the best performing vocabulary size is smaller than the number of keypoints. In other words, if some keypoints are similar enough to each other, they should be represented by one single visual word instead of many different visual words separately in order to obtain the best performance. From this perspective, we see that with any descriptor of limited dimension, all the possible keypoints can be represented by a limited number of visual words. Since this vocabulary of limited size covers the whole feature space, it can be used as a universal visual vocabulary. In Sections 2 and 3, we empirically validate this hypothesis.

The existence of a universal and compact visual vocabulary can be understood as follows. In the text domain, the number of words with semantic meanings is limited and the vocabulary is of a limited size. However, this vocabulary of limited size is able to deliver numerous meanings which we intend to express. The reason, in our opinion, lies in that each word in the vocabulary delivers not just one basic meaning, but also some other meanings similar enough to the basic one. In other words, each word in the vocabulary covers not just one point in the space of semantic meanings, but also the neighborhood of the point. In this way, a vocabulary of limited size is able to cover the whole space of semantic meanings and be used as a universal vocabulary. From this perspective, it is easy to understand why a compact visual vocabulary is able to represent all the image patterns with semantic meanings accurately and is universal across image datasets.

The observation in [3] that a larger size than the optimal may result in performance loss is explained in the following. Theoretically, visual vocabularies of optimal and larger sizes are all able to describe all the possible image patterns and their performances should be identical. However, with the increase of vocabulary sizes, the number of keypoints represented by one visual word tends to decrease. As a result, the calculation of a visual word as the centroid of a group of keypoints is more likely to be influenced by noise, which, in turn, harms the performance of visual vocabularies. Another interesting observation is that the optimal vocabularies from different datasets are of different sizes and they all can be used as the universal one. Our explanation is that when the vocabulary size is large enough, the relatively small variance in vocabulary sizes has little influence on the representation of image content, as illustrated in Table 1.

Based on the above analysis and the experimental results in Sections 2 and 3, we believe we have found out something related to the mechanism of bag-of-visual-words. In the next step, we plan to deepen our work by borrowing ideas and methods from related domains, for example, data-driven approaches [30–35] and psychological research.

5. Conclusion

Bag-of-visual-words is an important image representation and has been widely used in computer vision and pattern recognition problems. While much works have been published surrounding this representation, almost all existing works are based on the implicit assumption that a good visual vocabulary is data dependent. The problem of building a universal visual vocabulary of limited size is rarely touched.

Based on previous works on classification performance with respect to the vocabulary sizes, we arrive at the hypothesis that when features are similar enough, they should be represented by one visual word in order to obtain the best classification results. This further indicates that the whole feature space can be represented by a limited number of visual words, which constitute a universal visual vocabulary of limited size.

Starting from this hypothesis, we proposed to use a similarity threshold based clustering method to calculate the optimal vocabulary size for a given dataset. The optimal vocabulary size is then used to generate the optimal visual vocabulary for the dataset. In the experiments, we found that the three optimal vocabularies of limited sizes (several thousands) built from three datasets separately are very similar to each other, and any of them can be used to generate the best or near-best performance with all three datasets. This encouraging result indicates that with more datasets involved, it is really feasible to obtain a universal and compact visual vocabulary to be used in any datasets to generate (near-)best performance.

We analyzed the reasons behind the existence of a universal and compact visual vocabulary and other related phenomena observed in experiments. Based on the experiments and our analysis, we believe that we have found out something underlying the behavior of the bag-of-visual-words representation. Since in text domain a universal vocabulary is usually used with different datasets, our work narrows the gap between bag-of-visual-words and bag-of-words. This may lead to new approaches to explore the potential of the bag-of-visual-words representation and other image representation methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant no. 61304102), the Natural Science Foundation of Liaoning Province of China (Grant no. 2013020002), and the Scientific Research Fund of Liaoning Provincial Education Department of China (Grant nos. L2012400 and L2012397).

References

- [1] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1470–1477, October 2003.
- [2] S. Lazebnik, C. Schmid, and J. Ponce, "A maximum entropy framework for part-based texture and object recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 832–838, October 2005.
- [3] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the International workshop on Multimedia Information Retrieval*, pp. 197–206, September 2007.
- [4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2169–2178, June 2006.
- [5] M. Marszałek and C. Schmid, "Spatial weighting for bag-of-features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2118–2125, June 2006.
- [6] V. Viitaniemi and J. Laaksonen, "Spatial extensions to bag of visual words," in *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '09)*, pp. 280–287, July 2009.
- [7] D. Nistér and H. Stewénus, "Scalable recognition with a vocabulary tree," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 2161–2168, June 2006.
- [8] H. Cai, F. Yan, and K. Mikolajczyk, "Learning weights for codebook in image classification and retrieval," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2320–2327, June 2010.
- [9] T. Li, T. Mei, and S. K. In, "Learning optimal compact codebook for efficient object categorization," in *Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV '08)*, pp. 1–6, January 2008.
- [10] P. K. Mallapragada, R. Jin, and A. K. Jain, "Online visual vocabulary pruning using pairwise constraints," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, vol. 2, pp. 3073–3080, June 2010.
- [11] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: an in-depth study," Tech. Rep., INRIA, 2003.
- [12] W. Zhao, Y. Jiang, and C. Ngo, "Keyframe retrieval by keypoints: can point-to-point matching help?" in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 72–81, 2006.
- [13] J. Hou, Z. S. Feng, Y. Yang, and N. M. Qi, "Towards a universal and limited visual vocabulary," in *Proceedings of the International Symposium on Visual Computing*, pp. 414–424, 2011.
- [14] C. X. Ries, S. Romberg, and R. Lienhart, "Towards universal visual vocabularies," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '10)*, pp. 1067–1072, July 2010.
- [15] J. Hou, W. X. Liu, X. E. Q. Xia, and N. M. Qi, "An experimental study on the universality of visual vocabularies," *Journal of Visual Communication and Image Representation*, vol. 24, pp. 1204–1211, 2013.
- [16] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1800–1807, October 2005.
- [17] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1243–1256, 2008.
- [18] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, December 2008.

- [19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [20] J. Hou, N. M. Qi, and J. X. Kang, "Image matching based on representative local descriptors," in *Proceedings of the International Conference on Multimedia Modeling*, pp. 303–313, 2010.
- [21] J. Hou, J. X. Kang, and N. M. Qi, "On vocabulary size in bag-of-visual-words representation," in *Proceedings of the Pacific-Rim Conference on Multimedia*, pp. 414–424, 2010.
- [22] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories," in *Proceedings of the Workshop on Generative-Model Based Vision (CVPR '04)*, p. 178, 2004.
- [23] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [24] F.-F. Li and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 1, pp. 524–531, June 2005.
- [25] L. L. Jia and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV '07)*, pp. 1–8, October 2007.
- [26] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, Caltech, 2007.
- [27] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, pp. 1447–1454, June 2006.
- [28] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from National University of Singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '09)*, pp. 1–9, July 2009.
- [29] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," in *Proceedings of the European Conference on Computer Vision*, pp. 71–84, 2004.
- [30] S. Yin, X. Yang, and H. R. Karimi, "Data-driven adaptive observer for fault diagnosis," *Mathematical Problems in Engineering*, vol. 2012, Article ID 832836, 21 pages, 2012.
- [31] S. Yin, S. Ding, A. Haghani, H. Hao, and P. Zhang, "A comparison study of basic datadriven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process," *Journal of Process Control*, vol. 22, pp. 1567–1581, 2012.
- [32] P. Flandrin and P. Goncalves, "Empirical mode decompositions as data-driven wavelet-like expansions," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, pp. 1–20, 2004.
- [33] S. Yin, S. Ding, A. Haghani, and H. Hao, "Data-driven monitoring for stochastic systems and its application on batch process," *International Journal of Systems Science*, vol. 44, pp. 1366–1376, 2013.
- [34] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 1092–1099, November 2011.
- [35] S. Yin, H. Luo, and S. Ding, "Real-time implementation of fault-tolerant control systems with performance optimization," *IEEE Transactions on Industrial Electronics*, 2013.