Image Reconstruction from Bag-of-Visual-Words

Hiroharu Kato and Tatsuya Harada
The University of Tokyo
{kato, harada}@mi.t.u-tokyo.ac.jp

Abstract

The objective of this study is to reconstruct images from Bag-of-Visual-Words (BoVW), which is the de facto standard feature for image retrieval and recognition. BoVW is defined here as a histogram of quantized descriptors extracted densely on a regular grid at a single scale. Despite its wide use, no report describes reconstruction of the original image of a BoVW. This task is challenging for two reasons: 1) BoVW includes quantization errors when local descriptors are assigned to visual words. 2) BoVW lacks spatial information of local descriptors when we count the occurrence of visual words. To tackle this difficult task, we use a large-scale image database to estimate the spatial arrangement of local descriptors. Then this task creates a jigsaw puzzle problem with adjacency and global location costs of visual words. Solving this optimization problem is also challenging because it is known as an NP-Hard problem. We propose a heuristic but efficient method to optimize it. To underscore the effectiveness of our method, we apply it to BoVWs extracted from about 100 different categories and demonstrate that it can reconstruct the original images, although the image features lack spatial information and include quantization errors.

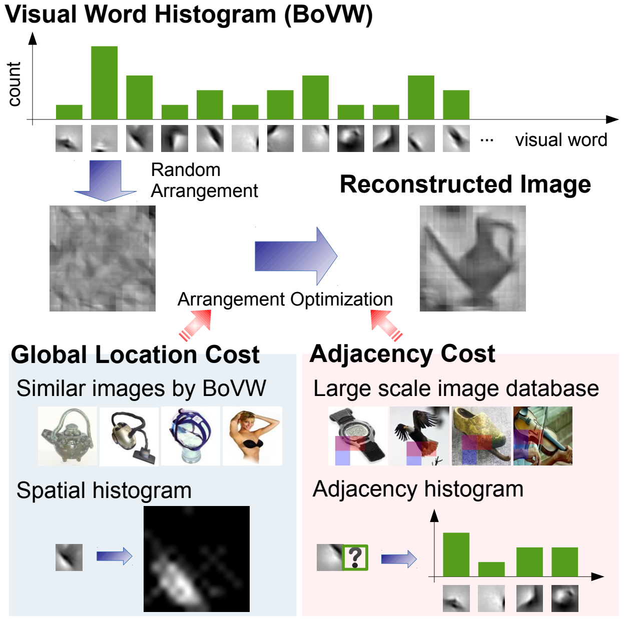


Figure 1. Visual illustration of our reconstruction method. The spatial arrangement of visual words is optimized. Then each visual word is converted to an image patch. An external image database is used for optimization.

1. Introduction

Image reconstruction from image features is attracting the interest of researchers in the computer vision community [8, 23, 25]. Image features are usually extracted from images through deep nonlinear transformations (e.g., local description and coding followed by spatial pooling). For that reason, it is not straightforward to estimate the original image. Although it might be possible to estimate the original image using an accurate image retrieval method and a large-scale image dataset, direct image reconstruction from a feature is expected to provide numerous benefits. For example, by observing the reconstructed images, one can understand intuitively what characteristics the image feature has, which helps us to examine the behavior of visual recognition or retrieval systems, and which suggests ways to improve their performance.

Bag-of-Visual-Words (BoVW) [6, 22] is the de facto standard of image features for retrieval and recognition. Many variants of BoVW have been proposed in the past (e.g., [17, 24]). In this work, we define BoVW as a histogram of quantized descriptors extracted densely on a regular grid at a single scale. Dense sampling with fixed grid spacing, descriptor scale, and orientation is a general setup of generic object recognition [3]. In the pipeline of BoVW, local descriptors are first quantized and assigned to visual words in the dictionary. Next, they are counted to form a histogram, which is treated as an image feature.

As described above, several methods have been proposed to reconstruct the original image from its features [8, 23, 25]. However, despite the importance of BoVW, no report in the literature describes reconstruction of the original image of a BoVW. Two hurdles must be overcome with this task: 1) Quantization errors occur in vector quantization of local descriptors. 2) Spatial information of local descriptors is ignored when the occurrence of visual words is counted to generate histogram-based features.

Can we reconstruct an original image from BoVW? To address this challenging problem, we use a large-scale image database to recover the spatial arrangement of local descriptors. Figure 1 presents a visual illustration of our image reconstruction method. We first search similar images by BoVW and calculate global location costs of visual words. We also estimate the adjacency cost of local descriptors in the database. By considering both the global location cost and local adjacency cost of descriptors, the task results in an optimization problem of a jigsaw puzzle. Solving it is known as an NP-Hard problem. We propose a heuristic but efficient method to optimize it.

Our contributions are the following. 1) This is the first work to tackle the problem of image reconstruction from BoVW, which would engender the improvement of numerous computer vision systems. 2) The bottleneck of that problem is recovery of the spatial arrangement of local descriptors. We define it as an objective problem and propose a means to determine parameters that use a large-scale image database. 3) We show the relations among our objective function, jigsaw puzzle solvers, and the Quadratic Assignment Problem. We propose an efficient optimization algorithm based on it.

2. Related work

Few reports have described image reconstruction from image features. We present a brief review in this section.

A study by Weinzaepfel *et al*. [25] was the first work to tackle this problem. They used SIFT descriptors and their geometry information: the location, orientation, and size of the image patch from which the descriptor was extracted. For each descriptor, their method retrieves the nearest descriptor in a large-scale image database using nearest neighbor search. It then rotates and rescales the corresponding image patch. Finally they are blended into an image using Poisson blending [20].

d'Angelo *et al*. [8] converted BREAF descriptors [2] and FREAK descriptors [1] into image patches. They analytically constructed an image patch the descriptor of which is identical to the input. Consequently, their method requires no external image database. Moreover, it can do conversion instantly.

Vondrick *et al*. [23] inverted the HOG feature [7]. They proposed four algorithms and concluded that an approach based on learning of a pair dictionary of features and their corresponding images is effective. Their method is sufficiently fast to invert features on the spot. It is applicable to arbitrary features in principle.

As stated before, no report in the relevant literature has described a method to reconstruct images from BoVW, which lacks geometry information of local descriptors.

3. Reconstruction method

In this section, we propose a method to reconstruct images from BoVW. Evaluation metrics for reconstruction and re-arrangement are also described.

3.1. Problem settings

From a BoVW, we can ascertain the extent to which the visual words are contained in the original image. However we cannot know their geometry information. If geometry information is obtained somehow, the remaining task is to reconstruct an image from its quantized local descriptors and their geometry information. This is potentially solvable using the findings of previous reports.

Two strategies are available to extract local descriptors. One is sparse sampling using keypoint detectors. The other is dense sampling where descriptors are extracted densely on a regular grid with fixed scale. Because the latter is more advantageous for image retrieval [13] and recognition [19], we assume that dense sampling is used. Additionally, we assume that descriptors are extracted at a single scale for simplicity. We also assume that adjacent patches can be overlapped. It is a common setting for many state-of-the-art image recognition pipelines, and overlapping local patches generally produce better results, even in deep convolutional networks [15].

In this work, all information necessary to extract BoVWs from images is assumed to be available. In other words, we know the dictionary of visual words, the spacing of the grid of dense sampling, the size of an image patch for local description, and the size of images to be reconstructed. These are ordinarily available for administrators of computer vision systems.

From these assumptions, geometry information that should be recovered turns out to be only the spatial arrangement of quantized local descriptors. Therefore, the problem can be decomposed into two subproblems: to recover an arrangement of visual words and to reconstruct an image from the recovered arrangement. The former resembles jigsaw puzzle solvers [4, 5, 12, 21] that have been widely developed recently. The relation between our work and theirs is described in Section 3.5.

3.2. Objective function

In dense sampling, local descriptors are extracted at grid points. The problem here is to rearrange the extracted and quantized local descriptors, which lack spatial information, in a grid. Concretely, n visual words in an image will be assigned at n grid points. We represent an assignment by a permutation matrix x. If the i th visual word in BoVW is assigned to k th location, then $x_{ik} = 1$, otherwise $x_{ik} = 0$. x must satisfy the following constraints.

$$\textstyle\sum_{i=1}^{n} x_{ik} = 1 \quad (1 \leq k \leq n)\,. \tag{1}$$

$$\textstyle\sum_{k=1}^{n} x_{ik} = 1 \quad (1 \leq i \leq n)\,. \tag{2}$$

$$x_{ik} \in \{0, 1\} \quad (1 \leq i, k \leq n)\,. \tag{3}$$

To rearrange local descriptors appropriately, we use two strategies: 1) satisfying co-occurrence relations of neighboring local descriptors, and 2) considering a prior absolute location of each local descriptor. In the former strategy, it is similar to solving jigsaw puzzles considering the compatibility of edges, shapes, and colors in adjacent pieces. In the latter, it is similar to solving them assigning pieces of a sky to the upper part and pieces of a ground to the bottom part.

We define the cost of each approach as C^a and C^l, and also define total cost as a weighted sum of both costs. Concretely, we propose $C^a = \sum_{i,j,k,l=1}^{n} C^a_{ijkl} x_{ik} x_{jl}$. It is called adjacency cost and a cost to assign i, j th visual word in BoVW to k, l th location in the image. Furthermore, we define $C^l = \sum_{i,k=1}^{n} C^l_{ik} x_{ik}$. It is called global location cost and a cost to assign i th visual word to l th location. We define C^a and C^l as a frequency in an image database in analogy with language models in Natural Language Processing. We smooth, normalize, and take a negative logarithm of the frequencies. Details are described in the following subsections.

Introducing λ as a weighting parameter, the proposed optimization problem is summarized as follows.

$$\begin{aligned} \min \quad & \lambda \textstyle\sum_{i,j,k,l=1}^{n} C^a_{ijkl} x_{ik} x_{jl} \\ & +(1-\lambda) \textstyle\sum_{i,k=1}^{n} C^l_{ik} x_{ik}. \end{aligned} \tag{4}$$

$$\text{s.t.} \quad \text{Equation } 1,\ 2,\ 3. \tag{5}$$

3.3. Adjacency cost

Adjacency cost gives a reconstructed image consistent edges and shapes. Although it is not difficult to ascertain whether two raw image patches are compatible or not, it is problematic to measure the compatibility of two visual words because quantized local descriptors lack details of their corresponding image patches. Here we assume an adjacency of visual words in BoVW to be reconstructed is the same as that in large scale image database. From this perspective, we propose a means to obtain C^a as Algorithm 1. The lower right part of Figure 1 is an illustration of the algorithm.

Adjacency cost is defined as the negative logarithm of the normalized histogram of co-occurrences of pairs of visual words in a neighboring region. For all possible pairs of visual words and adjacent patterns, a large-scale image database is scanned to count the occurrences of the pair and the adjacent pattern. We discard the absolute positions of the pair and handle only the relative position. We also ignore pairs which are not in m-neighbor distance. For this study, we use $m = 48$, which is a 7 × 7 area centered at an element.

Algorithm 1 Determination of adjacency cost.

```
Input: image database, distance parameter $m$
Output: adjacency cost $C^a$
  initialize $C^a$ with zeros
  $n \leftarrow$number of positions in an image
  for each image in an image database do
    for each position $k$ in the image do
      for each position $l$ in the image do
        $w_k \leftarrow$ visual word extracted at $k$ in the image
        $i \leftarrow$ visual word number of $w_k$ in the dictionary
        $w_l \leftarrow$ visual word extracted at $l$ in the image
        $j \leftarrow$ visual word number of $w_l$ in the dictionary
        if $k$ is within $m$-neighbor to $l$ then
          $\boldsymbol{d} \leftarrow (x_k - x_l, y_k - y_l)$
          $C^{a'}_{ij\boldsymbol{d}} \leftarrow C^{a'}_{ij\boldsymbol{d}} + 1$
        end if
      end for
    end for
  end for
  $C^a \leftarrow C^a + 1$
  for all $i, k, l$ such that $1 \leq i, k, l \leq n$ do
    $\boldsymbol{d} \leftarrow (x_k - x_l, y_k - y_l)$
    normalize $C^a_{ij\boldsymbol{d}}$ such that $\sum_j C^a_{ij\boldsymbol{d}} = 1$
  end for
  for all $i, j, k, l$ such that $1 \leq i, j, k, l \leq n$ do
    $\boldsymbol{d} \leftarrow (x_k - x_l, y_k - y_l)$
    $C^a_{ijkl} \leftarrow -\log\left(C^a_{ij\boldsymbol{d}}\right)$
  end for
  return $C^a$
```

3.4. Global location cost

Global location costs make a reconstructed image globally feasible. The rough shape of the image to be reconstructed might resemble those of images similar to it. Fortunately, we can easily obtain similar images using BoVW because it is extremely useful for retrieval. From this perspective, we propose a means to obtain C^l as Algorithm 2. The lower left part of Figure 1 presents a visual illustration of the algorithm.

Global location cost is defined as the negative logarithm of the normalized histogram of the occurrence of a certain visual word at a certain location. To compute the histogram, we retrieve a hundred images using the nearest neighbor search with input BoVW. For all visual words and locations, we scan them to construct a histogram of a visual word and location.

Algorithm 2 Determination of the global location cost.

```
Input: similar images
Output: global location cost $C^l$
  initialize $C^l$ with zeros
  for each image in similar images do
    for each location $k$ in the image do
      $w \leftarrow$ visual word extracted at $k$ in the image
      $i \leftarrow$ visual word number of $w$ in the dictionary
      $C^l_{ik} \leftarrow C^l_{ik} + 1$
    end for
  end for
  $C^l \leftarrow C^l + 1$
  for each visual word number $i$ in the dictionary do
    normalize $C^l_{ik}$ such that $\sum_k C^l_{ik} = 1$
  end for
  $C^l \leftarrow -\log\left(C^l\right)$
  return $C^l$
```

Algorithm 3 Optimization of Equation 4.

```
Input: $C^a$, $C^l$
Output: optimal arrangement
  $population \leftarrow$ randomly generated initial solutions
  while $\max(population) \neq \min(population)$ do
    $parents \leftarrow$ randomly selected pair in $population$
    $child \leftarrow$ generate a new child from $parents$
    optimize $child$ by Hill Climbing
    if $child < \max(population)$ then
      if $\mathrm{rand}(0,1) < p$ then
        one of most similar pairs in $population \leftarrow child$
      else
        $\mathrm{argmax}(population) \leftarrow child$
      end if
    end if
  end while
  return $\mathrm{argmin}(population)$
```

3.5. Optimization

Our objective function 4 is a generalized version of a string of jigsaw puzzle solvers [4, 5, 12, 21]. In these studies, C^a_{ijkl} has a nonzero value only where location k is a four-neighbor to location l. We consider pairs in further distance because they can be overlapped. Cho *et al.* included the global location cost [5]. However, most studies included only the adjacency cost [4, 12, 21]. For optimization, greedy algorithms [4, 12], belief propagation [5], and genetic algorithms [21] are used.

On jigsaw puzzle problems, adjacency cost is rather accurate, which makes these algorithms work well. In our work, however, adjacency costs are less reliable, which requires more sophisticated optimization algorithms. Here we demonstrate that our objective function results in a Quadratic Assignment Problem (QAP) and adopt an optimization method for QAP.

Lawler's generalized formulation [16] of QAP [14] is defined as follows.

$$\min \quad \sum_{i,j,k,l=1}^{n} c_{ijkl} x_{ik} x_{jl}. \tag{6}$$

$$\text{s.t.} \quad \text{Equation } 1,\ 2,\ 3. \tag{7}$$

Our function 4 can be transformed as follows.

$$\min \quad \sum_{i,j,k,l=1}^{n} \left(\lambda C^a_{ijkl} + \frac{1-\lambda}{n^2} C^l_{ik}\right) x_{ik} x_{jl}. \tag{8}$$

Comparing Equation 6 with Equation 8, our optimization problem results in QAP.

QAP is known as an NP-Hard problem and solved mainly by a hybrid of two meta-heuristics [18]. The results of extensive experiments [10] demonstrate that an approach of Genetic Algorithm with tabu search [9] is effective. In this work, we combine a Genetic Algorithm and a Hill Climbing algorithm because of computational complexity. Algorithm 3 shows our optimization algorithm. The way to modify a solution, generate a child from parents, and find the most similar pairs is the same as Drezner's method [9]. We set the population size to 100 and the ratio of the replacement pattern p to 0.2.

The computational complexity of tabu search is $O(n^3)$ in his setting, where n is the number of elements to be arranged. We do not use tabu search but Hill Climbing, which reduces it from $O(n^3)$ to $O(n^2)$. Here we assume that $C^a_{ijkl} = 0$, where k and l are not in m-neighbor distance. Then efficient implementation can reduce it from $O(n^2)$ to $O(m^2)$. Consequently, it is reduced from $O(n^3)$ to $O(m^2)$. In the experiment section, we set $n = 169$ and $m = 48$. Therefore, the computational complexity is $O(10^5)$ times smaller than the original one.

3.6. Image generation

Using the procedures described above, an optimal spatial arrangement of quantized local descriptors is obtainable. The remaining task is to generate an image from them.

A local descriptor can be converted to an image patch using an arbitrary existing method [8, 23, 25]. Here we use HOGgles [23], which is applicable to arbitrary features in principle. Generated patches are arranged into an image. Overlapping parts are simply averaged.

3.7. Evaluation metrics

Actually, measuring the similarity between a reconstructed image and the original one is not so easy. Appropriate image features and metrics are necessary for it, although selecting ones equivalent to human senses is an ultimate goal of Computer Vision. In this paper, we use a naive method as the mean squared error of raw pixels (DIFF) to

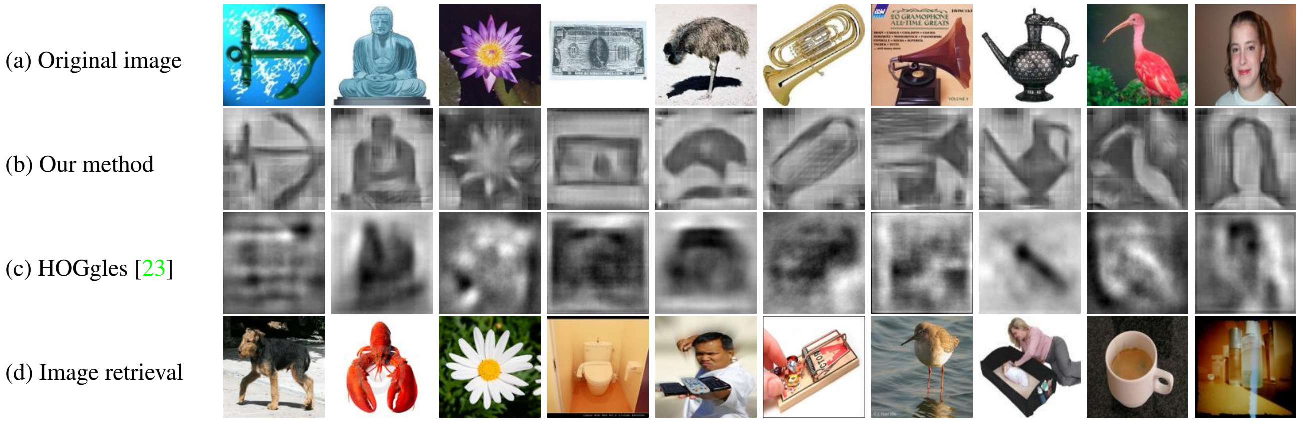


Figure 2. Examples of images obtained from BoVW.

eliminate the arbitrariness. For the reason that it can be influenced easily by a small parallel shift of images, we also compute it by shifting one image by ± 4 (DIFF4) pixels or ± 8 pixels (DIFF8), and select the minimum value.

Cho *et al*. [5] proposed performance metrics for jigsaw puzzle solvers. To evaluate the accuracy of descriptor rearrangement, we compute two metrics from their work. One is direct comparison (DC), which is the fraction of visual words that are assigned to the same location in the original image. The other is called neighbor comparison (NC), which is the fraction of pairs of visual words in four-neighbor distance, which are adjacent correctly.

3.8. Computational cost

The computational cost of reconstruction can be reduced by precomputations. Parameters of adjacent cost can be computed on ahead. In addition, a table of visual words at location l of image i can be made in advance. It enables computation of global location cost efficiently from similar images. These precomputations take several hours.

In our experiment, reconstruction of an image takes about a minute. The bottlenecks are image retrieval and optimization. The former would be accelerated drastically if an efficient approximate searcher were used. The latter is not much scalable with the number of descriptors in an image because we assigned weight to optimization accuracy. However, any fast but less accurate optimization method is applicable to our pipeline. Additionally, if Spatial Pyramid [17] is used, the computational cost of optimization is reduced considerably because the possible positions of each visual word are restricted.

4. Experiments

In this section, we present empirical evaluations of our proposed method. We compare it with HOGgles [23] and image retrieval. The effects of a parameter λ, descriptor quantization, and optimization methods are also examined.

We composed an image reconstruction dataset by extracting images from the Caltech 101 dataset [11], which contains 101 object categories. From each category, we randomly selected one image aspect ratio close to 1. Part of our dataset is shown in the first row in Figure 2. Our entire dataset and all reconstructed images in this section are available at our website[1].

Unless otherwise noted, all images are resized to 128×128 pixels, the size of the visual word dictionary (dimension of BoVW) is 5000, the kind of local descriptor is SIFT, the size of image patches for local description is 32×32 pixels, and the extraction step is 8 pixels. One million images from ILSVRC 2012 image classification task[2] are used as an image database.

4.1. Reconstruction from local descriptors and their geometry information

Our method decomposes the problem into two subproblems. One is recovery of geometry information of local descriptors. The other is reconstruction of an image from them. Here we evaluate only the latter, assuming that geometry information of local descriptors is known. Results presented in Figure 3 show that images can be reconstructed finely even if descriptors are quantized.

4.2. Reconstruction from BoVW

In this section, we apply our method to the image reconstruction dataset. Although no other work reported image reconstruction from BoVW, a method to visualize HOG feature (HOGgles) [23] is applicable to arbitrary features, in principle. Additionally, image retrieval by BoVW can be a means to ascertain image contents of BoVW because it is favorable for retrieval. We also show and compare images

[1] http://www.mi.t.u-tokyo.ac.jp/kato/cvpr2014.html
[2] http://www.image-net.org/challenges/LSVRC/2012/

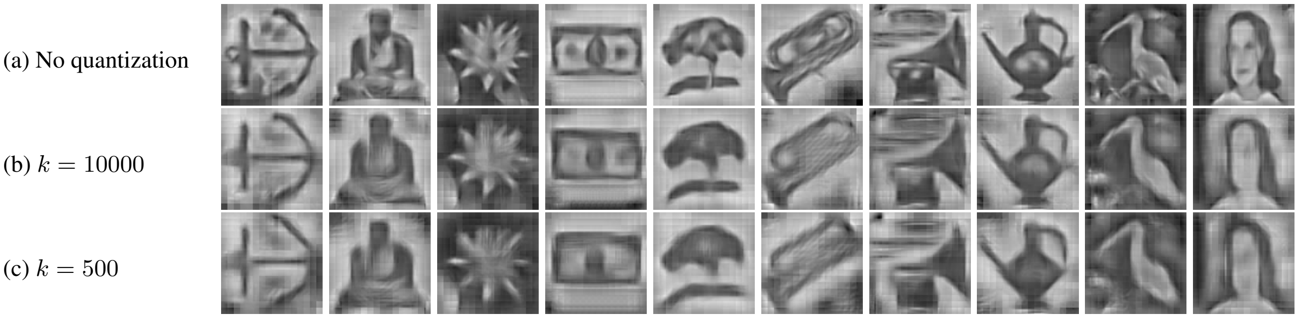


Figure 3. Examples of images reconstructed from local descriptors and their geometry information. k is the size of visual word dictionary. Images are understandable even if descriptors are quantized.

Table 1. Quantitative evaluation of the proposed method. HOGgles [23] and image retrieval (IR) were compared with our proposed method. DIFF, DIFF4, and DIFF8 are the mean squared errors between the original image and the obtained one. DIFFn shifts one image by $\pm n$ pixels and selects the minimum value. Values are averaged over 101 images. Smaller values are better.

	DIFF	DIFF4	DIFF8
Ours	**0.089**	**0.067**	**0.048**
HOGgles [23]	0.094	0.079	0.063
IR	0.111	0.090	0.071

Table 2. Winning rate of our method against other methods over 101 images. Methods are the same as those shown in Table 1.

	DIFF	DIFF4	DIFF8
Ours vs. HOGles	0.614	0.713	0.782
Ours vs. IR	0.713	0.812	0.841

obtained from these methods. We used nearest neighbor search to retrieve images.

Figure 2 portrays examples of images obtained using each method. They closely resemble results in Figure 3, which indicates that geometry information of descriptors can be well recovered. Images reconstructed using HOGgles are terribly blurred and are difficult to understand. Most images retrieved from the database are semantically different from the original images.

The success of recovery of location information has further meaning. Conventionally, it has been said that BoVW has no spatial information of each visual word. However, our results suggest that much location information remains potentially in BoVW.

To compare these methods quantitatively, we computed metrics of three kinds described in Section 3.7. Results are presented in Table 1 and Table 2. Our method is demonstrably superior to the other methods.

Figure 4 presents the best and worst results obtained using our method measured by NC. These results show that images that have simple shapes and backgrounds can be reconstructed well. However, those that have complicated texture and edges can not. The unreliability of adjacency cost is regarded as the cause.

4.3. Effect of parameter λ

Parameter λ adjusts the weights of adjacency cost and global location cost. To investigate the effect of this balancing, we reconstructed images for various λ. Figure 5 presents examples of reconstructed images. Figure 6 shows quantitative results.

When $\lambda = 0$, the adjacency of visual words is not considered, which results in heavily blurred images. When $\lambda = 1$, the similarity to similar images is not considered, which results in images with collapsed shapes. These results indicate that costs of two kinds in our objective function are working effectively. Quantitative results confirm that indication.

4.4. Effect of descriptor quantization

As shown in Section 4.1, descriptor quantization has little effect on converting local descriptors into image patches. However, it can affect their re-arrangement. When the visual word dictionary k becomes smaller, descriptors are quantized more strongly. Their details can be lost.

Figure 7 shows reconstructed images for various k. When k is small, shapes of images tend to be collapsed. As k becomes larger, reconstruction becomes stabler. Figure 8 shows that reconstruction error decreases as k increases.

To obtain good results, λ must be set larger when k becomes larger. As the vocabulary grows, quantization error decreases and accuracy of compatibility of adjacent elements increases, which eventually demands more weight on adjacency cost.

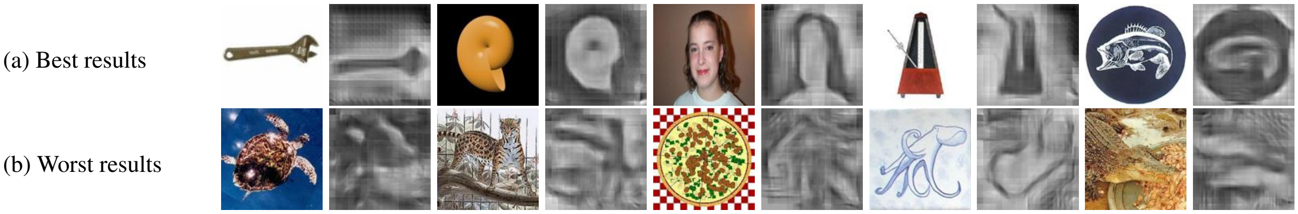


Figure 4. Best and the worst results. The top leftmost are the best. The bottom rightmost are the worst. Images are sorted by their neighbor comparison score. These results show that the accuracy of arrangement is dependent on the complexity of images.

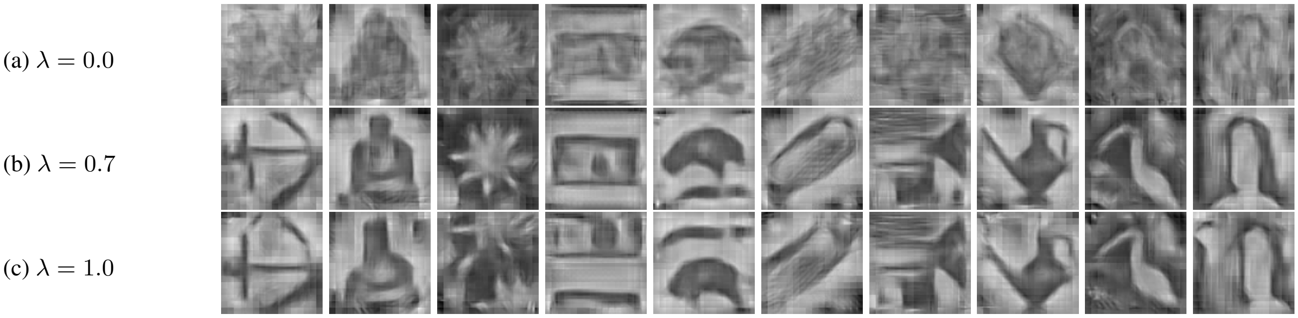


Figure 5. Examples of reconstructed images. λ is a parameter to balance two costs. These results illustrate the importance of balancing.

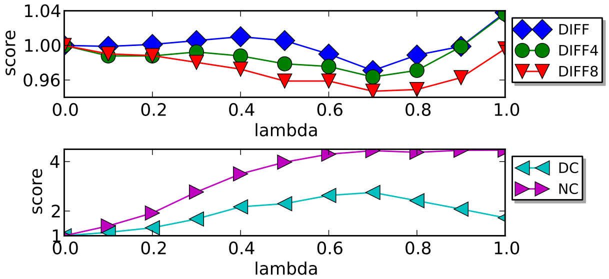


Figure 6. Quantitative evaluation of effect of parameter λ. Scores are averaged over 101 images and are divided by the score where $\lambda = 0$. For DIFF, DIFF4 and DIFF8, smaller values are better. For DC and NC, larger values are better.

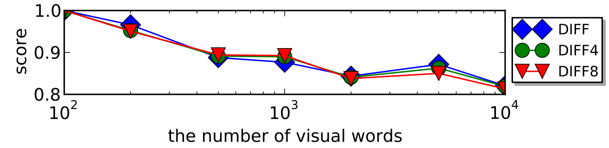


Figure 8. Relation between the size of visual word dictionary and reconstruction scores. Scores are averaged over the entire dataset and are divided by the score where $k = 100$.

4.5. Evaluation of optimization method

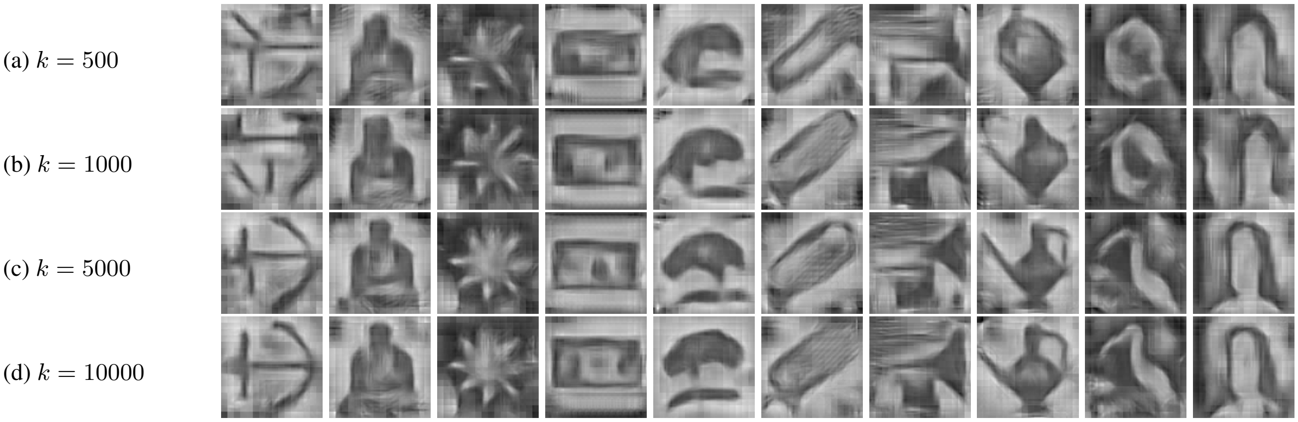
We compared our optimization method (GA+HC) with Hill Climbing (HC) and Simulated Annealing (SA). Table 3 shows the mean of optimized values and performance metrics over the entire dataset by each method. Results show that the optimization capability of our method is significantly better than that of the others.

Table 3. Comparison of optimization methods. Hill Climbing (HC), Simulated Annealing (SA), and a hybrid algorithm of Genetic Algorithm and Hill Climbing (GA+HC) were compared.

method	Eq. 4	DIFF	DC	NC
HC	769.1	0.099	0.051	0.303
SA	734.1	0.095	0.075	0.382
GA+HC	**719.0**	**0.089**	**0.195**	**0.459**

5. Conclusion

We presented a novel method to reconstruct an original image from its BoVW in which the spatial information of local descriptors disappears and quantization errors occur when local descriptors are assigned to the dictionary of visual words. The key techniques to succeed in the image reconstruction task are 1) modeling of adjacency and global location of visual words from a large-scale image database, 2) attributing the task to a jigsaw puzzle such as optimization problems with adjacency and global location costs, and 3) developing a heuristic but efficient optimization algorithm. In the experiments, we applied our method to recover 101 different object images from their BoVWs, and showed that our method can reconstruct the original images with a



(a) $k = 500$

(b) $k = 1000$

(c) $k = 5000$

(d) $k = 10000$

Figure 7. Examples of reconstructed images. The size of visual word dictionary k is varied. λ is manually adjusted to yield fine results: (a) $\lambda = 0.5$, (b) $\lambda = 0.6$, (c) $\lambda = 0.7$, and (d) $\lambda = 0.8$.

reasonable computational cost.

In this work, single scale sampling and hard assignment of local descriptors are presumed. These limitations can be relaxed if the cost function is extended. This extension is left as a subject for future work.

References

[1] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *CVPR*, 2012. 2

[2] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *ECCV*, 2010. 2

[3] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011. 1

[4] X. Chen, A. Jain, A. Gupta, and L. S. Davis. Piecing together the segmentation jigsaw using context. In *CVPR*, 2011. 2, 4

[5] T. S. Cho, S. Avidan, and W. T. Freeman. A probabilistic image jigsaw puzzle solver. In *CVPR*, 2010. 2, 4, 5

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[8] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *ICPR*, 2012. 1, 2, 4

[9] Z. Drezner. A new genetic algorithm for the quadratic assignment problem. *INFORMS Journal on Computing*, 15(3):320–330, 2003. 4

[10] Z. Drezner. Extensive experiments with hybrid genetic algorithms for the solution of the quadratic assignment problem. *Computers & Operations Research*, 35(3):717–736, 2008. 4

[11] 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. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. 5

[12] A. C. Gallagher. Jigsaw puzzles with pieces of unknown orientation. In *CVPR*, 2012. 2, 4

[13] A. Gordoa, J. A. Rodríguez-Serrano, F. Perronnin, and E. Valveny. Leveraging category-level labels for instance-level image retrieval. In *CVPR*, 2012. 2

[14] T. C. Koopmans and M. Beckmann. Assignment problems and the location of economic activities. *Econometrica*, 25(1):53–76, 1957. 4

[15] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[16] E. L. Lawler. The quadratic assignment problem. *Management Science*, 9(4):586–599, 1963. 4

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 5

[18] E. M. Loiola, N. M. M. De Abreu, P. O. Boaventura-Netto, P. Hahn, and T. Querido. A survey for the quadratic assignment problem. *European Journal of Operational Research*, 176(2):657–690, 2007. 4

[19] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006. 2

[20] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003. 2

[21] D. Sholomon, O. David, and N. S. Netanyahu. A genetic algorithm-based solver for very large jigsaw puzzles. In *CVPR*, 2013. 2, 4

[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1

[23] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013. 1, 2, 4, 5, 6

[24] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 1

[25] P. Weinzaepfel, H. Jégou, and P. Pérez. Reconstructing an image from its local descriptors. In *CVPR*, 2011. 1, 2, 4