Neurocomputing 119 (2013) 65–73

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Scene image retrieval via re-ranking semantic and packed dense interest points

Han Wang, Wei Liang*, Xinxiao Wu, Peng Teng

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Available online 17 January 2013

Keywords:
Content-based image retrieval
Scene image
Attribute
Packed dense interest points

ABSTRACT

In this paper, we propose a novel method for scene image retrieval in which the semantic meaning of an image and a new low-level feature are combined. The fluid nature of scene images makes learning semantics essential in our retrieval task. Compared to a general image, a scene image contains large regions of low contrast, which makes it difficult for a method to extract features that has good coverage of the entire image and assurance of relatively high repeatability. Given a scene image as a query, a collection of images is first retrieved by some search engines based on the images' semantic meanings. The candidate images are re-ranked by adapting an asymmetric piece-to-image matching scheme based on their visual similarities with the query image, using its visual signature consists of some packed dense interest points. Our method is evaluated on an Outdoor Scene Recognition (OSR) dataset and an NUS-WIDE dataset. It has demonstrated the improvements of our method over other conventional approaches.

1. Introduction

Retrieving images relevant to a keyword or an image query remains a very challenging task. Over many years, most image search engines have used keywords as queries and relied on their surrounding texts to search images. It is well known that textual descriptions fail to capture the essential characteristic. It is infeasible to pre-assign tags sufficiently to satisfy any future query a user may come up with. Content-based image retrieval (CBIR) [1–3], as we see today, is a technology that in principle helps to organize digital picture archives by their visual contents. It has become a major focus of computer vision and shown substantial progress. Scene image retrieval, viewed as a subfield of CBIR, has a potential to be applied in tourism advertising such as scenery spot searching and recommendation. For example, when we see a picture of a beautiful scenery spot that we are willing to be, we can easily find out the information about the site (or other sites with similar views) via a querying carried out by a CBIR search engine.

Conventional approaches [4,5] for general image retrieval tasks can be adopted in scene image retrieval as well. However, a major problem of conventional methods dealing with images lies in low-level feature extraction. As opposed to an "object" or a "textural" image, in a scene image, there is proposed to be relatively long distance between the observe and the fixated zone [6], which gives scene images a characteristic containing large regions with low contrast (uniform regions) such as sky, calm water, or flat road surface. Conventional methods mainly focus on the extracting visual signatures (mathematical descriptors of images) with high repeatability from the image, but consider less of interpreting it well. This is caused that, though the uniform regions in scene images contribute much to interpreting the scene, they are discarded by many visual signature extracting methods for retrieval. How to effectively extract more reasonable visual signatures of scene images is an open problem in scene image retrieval task.

Besides low-level similarity of images, the well-known and frustrating "semantic gap" between low-level visual cues and high-level user's intents remains, which makes it difficult to predict the behaviors of content-based scene searching systems. In other words, it is difficult to correlate low-level visual features with high-level semantic meanings that match the search intention of a user. For instance, two visually similar scene images may belong to different scene categories. In most cases, there is no clear definition as to which category a scene image belongs to. Therefore, it is essential to learn the semantic of scene images.

In this paper, we propose a new method for scene image retrieval in which an image's high-level and low-level features are combined to yield more reasonable retrieval results. Fig. 1 shows the framework. At the offline stage, semantic signatures of an image are first obtained with each signature predicting either the

* Corresponding author.
E-mail addresses: wanghan@bit.edu.cn (H. Wang), liangwei@bit.edu.cn (W. Liang), wuxinxiao@bit.edu.cn (X. Wu), tengpeng@bit.edu.cn (P. Teng).

0925-2312/$ - see front matter Crown Copyright © 2013 Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.06.048

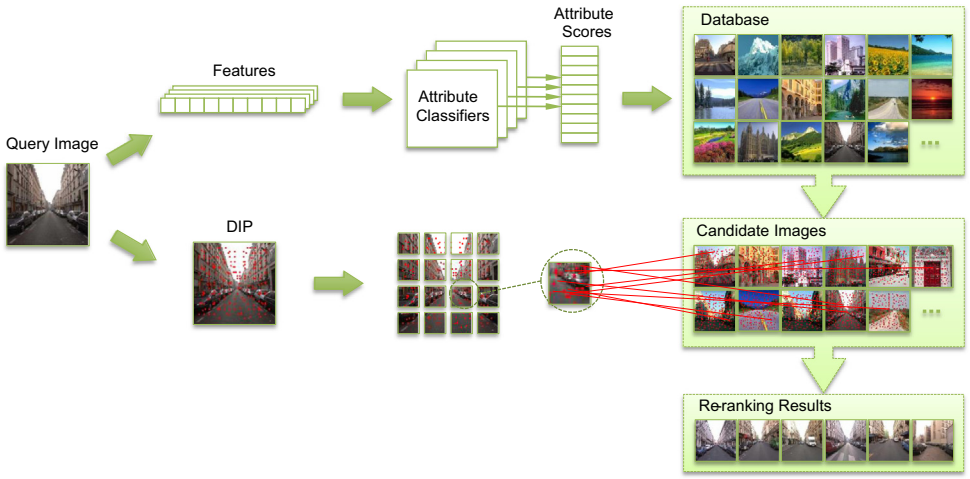

Fig. 1. Framework for scene image retrieval.

presence or the strength of a nameable semantic concept in the image. At the query time, candidate images are selected by their semantic similarities to the query image to ensure that the results meet the intention of the users. Then, these candidate images are re-ranked based on the low-level features, similarities to make the spatial layouts of query results consistent with that of the query image. Specifically, we use the attribute-based method to select candidate images with semantic similarities from the database, and develop a novel low-level feature called packed dense interest points as visual signature to re-rank the candidate images under asymmetric matching. We evaluate the method on an OSR dataset and an NUS-WIDE dataset. Experimental results demonstrate that our algorithm achieves superior image retrieval performance compared to conventional methods by discovering semantically meaningful descriptors of scene images.

2. Related work

We apply attributes of an image to find out semantically similar candidate images that most meet the user's intention, and consequently use a new low-level visual signature to re-rank images to obtain parallel layout images. We will review the research work on these two aspects in this section.

Works [7–9] were carried out to extract visual signatures. Features based on the local invariants were quite popular, for their efficient indexing and good discriminative power, owing to their compact representation of important image regions. Interest points (e.g. SIFT [10]) were just such features with high repeatability. They could be extracted reliably and often found again at similar location in other images of the same object or scene, which was a quite important property for visual signature. In scene image retrieval, with interest points one can choose an appropriate level of viewpoint and illumination invariance, and focus on the typical regions with high information content. However, interest points have limitations when they are performed on scene images containing large uniform regions.

As mentioned before, uniform regions have rather low contrasts, which result in too few or even no interest points found in these regions (see Fig. 2 for an instance). So, uniform regions cannot be interpreted well enough in visual signature extraction. Owing to the nature of CBIR, the missing part of interpretation will make negative influence on similarity measuring. In contrast to interest points, dense sampling on a regular grid is appropriate in dealing with scene images [11], since it focuses on all image regions and thus obtains a good coverage of the entire scene. Dense sampling is based on the idea that regions with different contrast contribute equally to the overall image representation. The comparative evaluation by Li and Perona [12] showed that dense features work well for scene classification. On the downside, the points obtained from dense sampling cannot reach the same level of repeatability as those obtained with interest points, unless this sampling scheme is performed extremely densely. This exhaustive sampling scheme will make the number of features quickly grow unacceptably large.

Recently, a hybrid scheme of interest points and dense sampling named dense interest points (DIPs) [13] is proposed. This method starts with the extraction of densely sampled patches are followed by optimizing their position and scale parameters locally. Their extraction starts from densely sampled patches and then optimize their position and scale parameters locally. DIPs are expected to be helpful in obtaining point features of a scene image with appropriate repeatability and good coverage for retrieval.

It is well known that similarity estimation plays a crucial role in image retrieval systems. Extensive research [14–16] has been dedicated to this topic for decades. In most cases, approximation methods still require computing pairwise candidate distortions prior to optimizing the assignment, which adds a significant computational overhead – $O(m^2n^2)$ – for images with m and n points. Recent progress in region-to-region matching includes a bag-of-regions scheme [17], graph matching approaches using region hierarchies [18], and multiple segmentation-based methods [19]. In contrast to these methods, we propose packed dense interest points [20] based on the DIPs as scene image visual signature and adopt asymmetric matching along with it. With such visual signature and via asymmetric matching, we can leverage the strength of spatial constraints on points while reducing the risk of getting stuck with poor segmentation only. However, this low-level feature matching scheme is time consuming when the database are extremely large. Furthermore, this method will result in finding two scene images of different semantic meanings. In order to obtain images meet the intention of a user, semantic learning becomes essential in content based image retrieval.

In order to obtain semantic meaning of a scene image, several investigations have been conducted [21–24] to build intermediate representations. The idea is to learn classifiers to predict the presence of various high-level semantic concepts from a lexicon (i.e. properties or location types), and then perform retrieval in

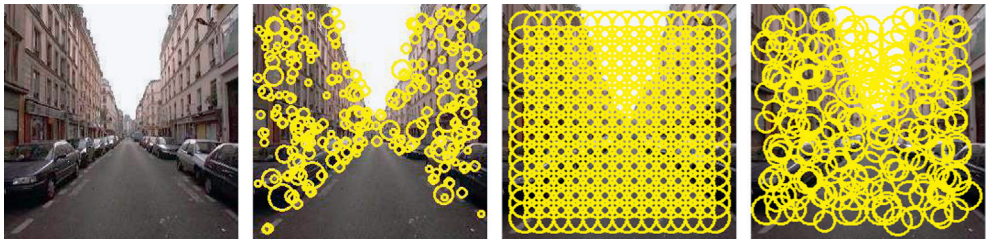

Fig. 2. Dense interest points (right most image) from an original image (left most). The second image represents interest points detection result, and the third image is dense sampling on a regular grid.

the space of those predicted concepts. Using attribute-based method, not only can one find semantic similar scene images but also can save a lot of time spent on useless matching. Furthermore, it is easy to add new images and scene attributes to the database, allowing for further scalability. To learn visual similarity metrics to capture the user's search intention, relevance feedback [25] is widely used. However, it requires efforts from the user side to select multiple relevant and irrelevant image examples, and often needs online training. For a web-scale commercial system, user feedback has to be limited to the minimum with no online training.

3. Retrieval with semantic similarity

We are given a set of images $\{i\}$ represented by $\boldsymbol{x_i} \in \mathbb{R}^n$, where $\boldsymbol{x_i}$ is the low-level feature vector represents the ith image, and a set of M attributes $A=\{a_m\}$, where a_m represents the mth attribute. At the semantic retrieval stage, in order to automatically obtain the semantic signatures of the images, classifiers of relative attributes are trained and stored offline. These relative attributes allow us to transform $\boldsymbol{x_i} \in \mathbb{R}^n$ to $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \in \mathbb{R}^M$, so that each image i is now represented by an M-dimensional vector $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$ that indicates its strength for all M attributes.

Here we use a Gaussian distribution to build a generative model for each category in semantic space. The mean $\boldsymbol{\mu}_c \in \mathbb{R}^M$ and covariance matrix $\sigma_c \in M \times M$ are estimated from the attribute-features of the training images belonging to category c. Consequently, we have $\mathcal{N}(\boldsymbol{\mu}_c,\sigma_c)$ for category c based on the relative attributes, where $c=1,\ldots,S$ with S indicates the number of categories. At the semantic query stage, once the user chooses a query image i, $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \in \mathbb{R}^M$ is computed as the semantic signature, and then assigned to the category c^* of the highest likelihood: $c^* = \arg\max_{c^* \in \{1,\ldots,S\}} P(\tilde{\boldsymbol{x}}_{\boldsymbol{i}} \mid \boldsymbol{\mu}_c,\sigma_c)$. All images in the dataset are ranked by their category c^* likelihood, and a set of top 100 candidate images are automatically selected. Our experiments show that adding semantic meanings of the images will increase the re-ranking accuracy, and meanwhile also increase the storage and reduce the online semantic retrieval efficiency because of the increased size of semantic signatures.

4. Re-ranking using packed features

After the semantic retrieval stage, the candidate images returned from the attribute indexing are initially ranked based on the similarities of their semantic signatures to those of the query image. Then these images are re-ranked according to their similarity with the query image. This way we can retrieval images with the parallel layout without exhaustively comparing every image in the dataset.

4.1. Packed dense interest points

Given a candidate image, we first detect its dense interest points and describe these points with SIFT [10]. Then we integrate point features with packed features with spatial relation among points to improve similarity matching.

4.2. Point features

For image representation, we use dense interest points feature extraction scheme, where interest points and dense sampling are combined to get the advantages of both worlds. It takes two steps to find the dense interest points. In the first step, we begin with densely sampled image patches on a regular grid. On a larger scale, a coarser grid is used. That is, the same grid is applied to multiple downscaled versions of the original image in a scale space pyramid, and Difference-of-Gaussian (DoG) scale pyramid [10] is adopted in this paper. In detail, on each patch, a local maximum of DoG is found by searching a larger area that extends beyond the direct neighboring pixels and computing patches over a grid with spacing of multiple scales (16 pixels $\times$ 16 pixels $\times$ 8 scale levels in our experiments), instead of checking for each pixel whether it is a local maxima in a neighborhood. Each maximum so obtained is retained as a dense interest point and further used to refine the patch position. We center the patch around this point and extract feature on the refined patch. To achieve better discriminative power, "strong features" (i.e. SIFT descriptors computed on a grid 16 pixels $\times$ 16 pixels $\times$ 8 scale levels) are used. To avoid bringing noise to features extracted on homogeneous image areas, we slightly modify the original SIFT feature normalization procedure. The vectors are normalized only if the norm of SIFT feature is larger than 1; otherwise, the original descriptor is kept without normalization.

4.3. Packed features

As mentioned in [9], to increase the discriminative power of a single local descriptor, the most direct approach is to increase the dimensionality of the descriptor or the region size in the image. For high dimensionality features, their repeatability may decrease due to their sensitivity to occlusion and other image variations, such as photometric and geometric changes. Therefore, we combine several point features within a region to form a packed feature to increase discriminative power of the local features. In our experiments, a 256×256 image is divided into 16 pieces with each piece corresponding to a 64×64 rectangular area (see Fig. 3). Dense interest points that fall inside the same piece are packed together as a packed feature. Let $\tilde{P}=\{\tilde{p}_i\}$ denote the dense interest points and $R=\{r_i\}$ denote the pieces apart from a query

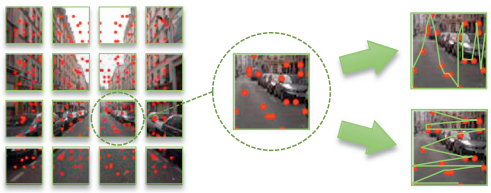

Fig. 3. Illustration of packed features.

image. Define the packed features of an image $I=\{P_i\}$ as

$$P_i=\{p_k|p_k=\tilde{p}_j,\tilde{p}_j \text{ in } r_i,\tilde{p}_j\in\tilde{P}\}, \tag{1}$$

where $\tilde{p}_j$ in r_i means that the jth point feature $\tilde{p}_j$ falls inside the ith piece r_i. The packed feature of the ith piece P_i is an ordered set in which elements are sorted by their spatial relations. In order to represent a 2D layout of the image by 1D sequences, each packed feature is arranged as two sequences of its constituent feature points. These two sequences are used to represent a piece: one sequence links points along the column and the other links those along the row. If there are more than one p_j in a patch with similar SIFT feature in a given range, the one with the largest magnitude is retained. Afterward, the number of features may decrease dramatically, since there are a lot of redundant features. Because a packed feature consists of multiple SIFT features, it allows partial matching of two groups of SIFT features. Thus is more discriminative than a single SIFT feature.

5. Asymmetric matching

This section shows how to measure the similarity of two images via efficient partial matching of packed features with weak geometric constraints in our scene image retrieval system. We first describe our piece-to-image mapping scheme and cost function in detail (Sections 4.1 and 4.2), and then calculate the matching score of the two images as the result of their similarity measuring (Section 4.3).

5.1. Piece-to-image mapping

In this step, each piece of the query image is mapped to a set of dense interest point descriptors. A packed feature is matched with the candidate image as follows. For each element in the packed feature, we first find a set of candidate matching points in the candidate image according to their Euclidean distances in the SIFT feature space. Next, we re-check the point sets and select the optimal correspondence by dynamic programming to minimize a cost function defined in Section 5.2. The cost function not only restricts the appearance difference between the corresponding points but also penalizes the large geometric distortion between the points and their neighboring points. This scheme can match a piece with any portion of the target image.

5.2. Cost function

We denote by $P_i=\{p_1,\ldots,p_{l_i}\}$ the packed features of a query image, where p_i represents the feature point extracted by the method discussed in Section 4.1 and l_i denotes the length of the ith piece's sequence. Let C_k represent the candidate set for the point p_k. We want to select the optimal matching sequence $Q_i=\{q_1,\ldots,q_{l_i}\}$ from candidate sets $C_1,\ldots,C_{l_i}$, where $q_i\in C_i, 1\le i\le l_i$. The cost function for the selection is given by

$$C(P,Q)=\sum_{k=1}^{l_i-1}\omega_g G(p_k,p_{k+1},q_k,q_{k+1})+\sum_{k=1}^{l_i}\omega_a A(p_k,q_k)$$
$$+\sum_{k=1}^{l_i}\omega_d D(z_i,q_k)+\sum_{k=1}^{l_i-1}\omega_o O(p_k,p_{k+1},q_k,q_{k+1}). \tag{2}$$

The cost function contains four terms that refer to both appearance and geometric consistency. The coefficients $\omega_g,\omega_a,\omega_o,\omega_d$ in the four sums weights showing the impacts of relevant terms. Some details of this function are as follows.

Geometric distortion $G(p_k,p_{k+1},q_k,q_{k+1})$: This term measures geometric disparity between two corresponding pairs (p_k,p_{k+1}) and (q_k,q_{k+1}):

$$G(p_k,p_{k+1},q_k,q_{k+1})=\|(p_k-p_{k+1})-(q_k-q_{k+1})\|_2, \tag{3}$$

where p_{k+1} is the point next to p_k in the packed feature sequence, q_k and q_{k+1} are corresponding matching points associated with p_k and p_{k+1}, respectively.

Appearance similarity $A(p_k,q_k)$: This term measures the appearance difference of the kth matching pair:

$$A(p_k,q_k)=\left(1+\exp\left(-\tau_a\left(\frac{1}{\mu_a}\|f_{p_k}-f_{q_k}\|_2-1\right)\right)\right)^{-1}, \tag{4}$$

where f_{p_k} and f_{q_k} are SIFT features extracted on the point p_k and q_k, respectively, and μ_a and τ_a are constants adjusting the shape of sigmoid function.

Displacement constraint $D(z_i,q_k)$: This term penalizes large displacement between the locations of a piece center and its corresponding points set:

$$D(z_i,q_k)=\begin{cases}\|z_i-q_k\|_2 & \text{if } \|z_i-q_k\|_2>t,\\ 0 & \text{otherwise,}\end{cases} \tag{5}$$

where z_i is a two-dimensional coordinate indicating the ith piece center position.

Ordering constraint term $O(p_k,p_{k+1},q_k,q_{k+1})$: This term penalizes pairs of correspondences when that violate the geometric ordering. Its value is 1 if the ordering of p_k and p_{k+1} is different from that of m_k and m_{k+1} (in either the horizontal or the vertical direction), and 0 otherwise. For example, if the point p_k is located left of the point p_{k+1}, its matching point m_k should be located left of m_{k+1}.

5.3. Similarity measurement

In this step, we assign a score of similarity for two images: query image I_1 and target image I_2. Similarity score for a single patch in query image is expressed as

$$S(I_1,I_2)=\frac{1}{\sqrt{LN}}\sum_{i=1}^{L}\omega_i s_a(P_i)s_g(P_i), \tag{6}$$

where the functions $s_a(\cdot)$ and $s_g(\cdot)$ score a piece's appearance and geometric similarity, L is the total number of pieces in the query image, N is the total number of points in the target image, and ω_i is a weight associated with the ith piece (will be defined below), respectively.

The appearance term in the cost function maps to the following score:

$$s_a(P_i)=1-\sum_{k=1}^{l_i}A(p_k,q_k). \tag{7}$$

The geometric consistency score measures the average distance between corresponding points in a pair and the rest of all matching pairs in the final optimal

$$s_g(P_i)=\frac{1}{n-1}\sum_{k=1}^{l_i}\left(\frac{1}{1+\exp\left(\tau_g\left(\frac{G(p_k,p_{k+1},q_k,q_{k+1})}{a_{ik}}-1\right)\right)}\right), \tag{8}$$

where $G(\cdot)$ is defined in the cost function, $\omega_i = |\{p_i\}|$ is a weight factor, and $k \in \{1,\ldots,n\}$ for each i. This term gives higher weight to the matched pieces with more common feature points, enforcing a weak spatial consistency. The weight is normalized by the total number of matched and unmatched features in P_i, so pieces with more features will produce more impact on the overall score.

6. Experiments

6.1. Datasets and implementation details

In our work, two image datasets (OSR [6] and NUS-WIDE [26]) are used to evaluate our proposed method.

OSR dataset: There are 2688 scene images splitted into eight categories in the dataset: coast, forest, highway, inside city, mountain, open country, street and tall building. Some examples are demonstrated in Fig. 4. We use 240 images for training (i.e. 30 images per category) and leave the rest as test set.

NUS-WIDE dataset: This dataset contains 269,648 Flickr images, which are divided into a training set (161,789 images) and a test set (107,859 images). It is manually labeled with 81 semantic concepts, covering a wide range of semantic topics from objects to scenes. The dataset we used in our experiments is a lite version of NUS-WIDE dataset (NUS-WIDE-SCENE) which covers 33 scene concepts with 34,926 images in total. We use half of these images (i.e. 17,463 images) for training and leave the rest (i.e. 17,463 images) for testing.

At the semantic-retrieval stage, according to the query image's semantic signature, the most relevant 100 images in the target dataset are selected as candidate images. At the re-ranking stage, for a query image, packed DIPs are used as its visual signature. First, point features are simply extracted according to the dense interest points extraction scheme, and about 1700 points are generated for a typical image size (256×256). Then, the image is divided into 4×4 pieces, each piece is represented by a packed dense interest points feature. For all the candidate images, we only extract dense interest point as their visual signature. After that, asymmetric piece-to-image matching is conducted for similarity measuring.

The initial candidate matching points set of an element in packed feature is one containing points with a SIFT descriptor distance within $1.25\mu_a$. We use approximate nearest-neighbor search [27] to find close feature points quickly. In the cost function, weight factors associated with the cost terms scale their relative impact and are empirically fixed as follows: $\omega_g = 1.0$, $\omega_a = 1.25$, $\omega_o = 1.5$, and $\omega_d = 4.0$. Here we randomly pick 10 images per category from the test set as query images, and leave the rest as target images. For each query, the scores of similarity computed by our matching method are ranked and the top 20 matches are returned as retrieval results.

Fig. 4. Examples of semantic retrieval on OSR.

6.2. Results and discussions

6.2.1. Discussion on computational cost and storage

Compared with the conventional image re-ranking scheme, our approach is much more efficient at the online stage, because the main computational cost of online image retrieval is on comparing visual features, and the lengths of semantic signatures are much shorter than those of low-level visual features. As a result of semantic selecting, the re-ranking image scale reduces dramatically. According to our experimental study, it takes less than half an hour to learn the semantic spaces of six attributes using a machine with Intel Core2 Q9400 2.66 GHz CPU. And for larger semantic space as given in NUS-WIDE dataset, we are able to process 33 attributes within 10 h. And the following re-ranking scheme takes about 3.5 h to compute similarity scores with the query image and the top 100 candidate images. With the fast growth of GPUs, which achieves hundreds of speedup than CPU, it is feasible to process the industrial scale queries. The extra storage of classifier and semantic signatures are comparable or even smaller than the storage of low-level features of images.

Table 1
Relative ordering of categories by attributes. The OSR dataset includes images from the following categories: coast (C), forest (F), highway (H), inside-city (I), mountain (M), open-country (O), street (S) and tall-building (T).

Attributes	Relative
Natural	$T < I \sim S < H < C \sim O \sim M \sim F$
Open	$T \sim F < I \sim S < M < H \sim C \sim O$
Perspective	$O < C < M \sim F < H < I < S < T$
Large-objects	$F < O < M < I \sim S < H \sim C < T$
Diagonal-plane	$F < O \sim M < C < I \sim S < H < T$
Close-depth	$C < M < O < T \sim I \sim S \sim H \sim F$

6.2.2. Evaluation on semantic ranking

We first examine the image retrieval performance by the semantic ranking step of our experiments. At the offline stage, for OSR dataset, we simply use the gist descriptor [6] as low-level features and use the relative attributes [28] as semantic signature of the images in the database. Table 1 illustrates the relative attribute assignment used in our experiments. For NUS-WIDE-SCENE dataset, the low-level features include 64-D color histogram, 144-D color auto-correlogram, 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments provided in [26]. We combine these five kinds of features directly by merging the five feature vectors, and SVM classifiers for each attribute are trained on these features. The 33 concepts defined in [26] are used as the semantic space of NUS-WIDE-SCENE dataset and the decision values are computed as relative attribute. Figs. 5 and 6 show the average precision (AP) and average recall (AR) on OSR dataset and NUS-WIDE-SCENE dataset, respectively. The semantic ranking reaches the average precision of 53.4% on the NUS-WIDE dataset and 72.7% on the OSR dataset. Fig. 4 shows top 9 candidate's images of some example queries on OSR dataset after semantic ranking. As shown in Fig. 4, the candidate images are semantic related, but the spatial layout of these images vary greatly.

6.2.3. Evaluation on re-ranking scheme

We evaluate our re-ranking scheme on the candidate images obtained from semantic retrieval step. Figs. 7 and 8 show average precision (AP) and average recall (AR) on both datasets. Our method

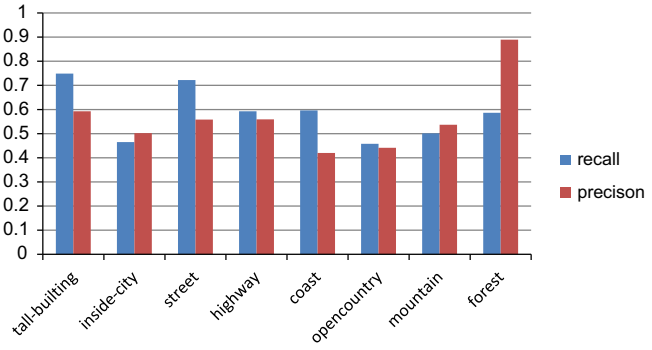

Fig. 5. Semantic retrieval performance on OSR.

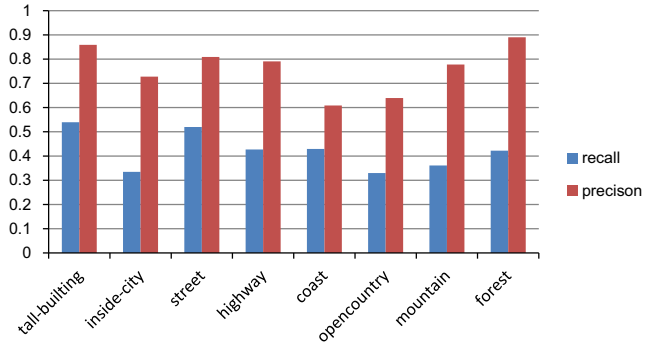

Fig. 7. Re-ranking with asymmetric matching based on the packed DIPs on OSR.

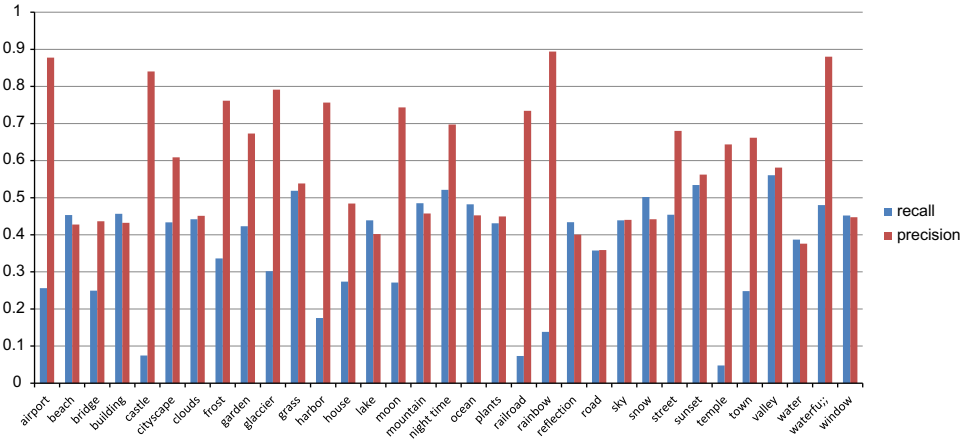

Fig. 6. Semantic retrieval performance on NUS-WIDE-SCENE.

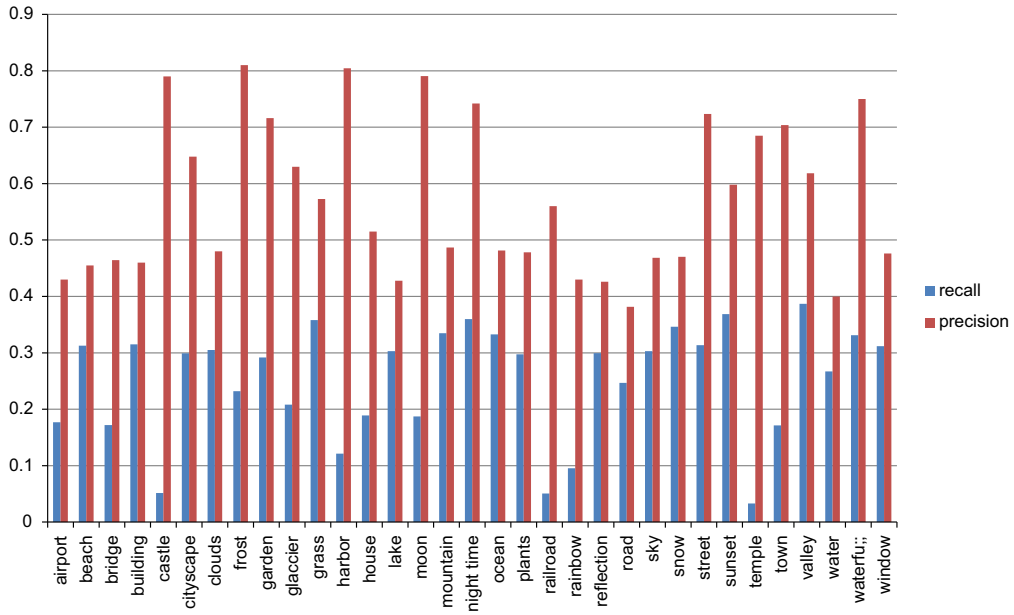

Fig. 8. Re-ranking with asymmetric matching based on the packed DIPs on NUS-WIDE-SCENE.

Fig. 9. Examples of retrieval results on the OSR.

Table 2
Comparison of our method to other image matching algorithms on OSR dataset.

Method	mAP (%)
SPM [14]	72.2
R-to-I [16]	59.2
Cluster sampling [29]	63.5
Packed DIPs [20]	67.4
Ours	**77.4**

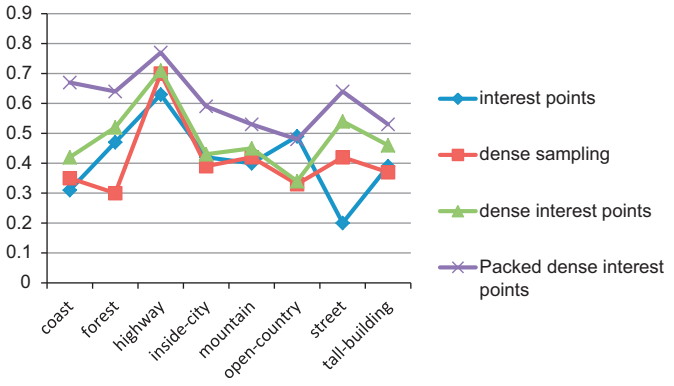

Fig. 10. Retrieval performances of different patch-based features.

reaches the average precision of 57.2% on the NUS-WIDE dataset and 77.4% on the OSR dataset (Table 2). When the re-ranking scheme is integrated, the average recall degrades while the average precision increases. Intuitively, owing to the constrain on parallel layout, and the images have different layouts in same categories will be excluded from the candidate images.

Next, the retrieval performance of our piece-to-image matching method with packed DIPs is compared with other image matching algorithms on the retrieval performance. All the related methods employ the nearest-neighbor for classification and rank the images based on the similarity scores. The retrieval accuracy is measured with the mean average precision (mAP) which is the mean precision of all the queries. Owing to the packed DIPs with the whole image information, our method achieves higher recognition accuracy than the methods [14,29], which also use the related pairwise geometric constraints. This is because our packed DIPs obtain the whole image information and attribute-base selection gets images which have the same visual concepts.

We show some image retrieval results with our framework in Fig. 9. The left-most image in each row is a query and the following nine are images returned from the database and ranked by their similarity scores. As shown in Fig. 9, some failure cases happen (e.g. building and mountain), for they have very similar geometric configurations which make it difficult to distinguish those images intuitively (Table 2).

6.2.4. *Evaluation on low-level features*

To evaluate the packed DIPs' performance in the context of retrieval task, the packed DIP feature is compared with some standard patch-based features [30]: standard interest points, dense sampling on the same regular grid, and dense interest points [13]. We conduct the experiment on the OSR dataset. We randomly pick 10 scene images per category as query images and calculate average precision for each category. From the comparison results shown in Fig. 10, we have the following observation: (1) For DIPs and packed DIPs, since our feature organization method deals with low-contrast areas as well as spatial relations of feature points, our packed features perform better on all categories. (2) DIPs outperform interest points in most instances. The reason is that DIPs can detect informative points appearing in uniform regions while the standard interest points cannot. However, for open-country, DIPs show degraded performance. That is because open country images have smaller uniform regions, so that the interest points can provide sufficient coverage of the image and point features with higher repeatability than DIPs. (3) Packed DIPs outperform dense sampling by extracting points which are more remarkable than those extracted by dense sampling. (4) Additionally, packed DIPs may overcome the negative effect from shortage of DIPs, such as the case in open-country, due to the asymmetric piece-to-image matching which is effective for imposing geometric constraints as mentioned in [16].

7. Conclusion

We have described a new approach searching for images from given scene image datasets, and introduced a scene image description called packed dense interest points. In the first step of search, semantic retrieval returns a set of candidate images that reflect the intention of the user. The second step re-ranks asymmetric matching based on the dense packed interest point that serve as constraints on the candidate images. As observing in the experiments, packed DIPs are a flexible representation with several desirable properties. First, they have the advantages of DIPs in dealing with uniform regions in scene images. Second, packing of the features allows some simple and robust geometric constraints to be imposed at the patch level, making them more discriminative than individual point features. Finally, under the introduced piece-to-image matching strategy, each piece can match points anywhere in the candidate images. Moreover, we show the power of combining high-level and low-level features to create a flexible architecture without sacrificing retrieval accuracy. In the future, we will incorporate global features into the framework to deal with cases where local features are insufficient to the task. We will also investigate on how to automatically define attributes from a training dataset with more discriminative and descriptive power.

References

[1] R. Datta, D. Joshi, J. Li, J. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Comput. Surv. CSUR 40 (2) (2008) 51.
[2] S. Siersdorfer, J. San Pedro, M. Sanderson, Automatic video tagging using content redundancy, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 395–402.
[3] X. Wang, L. Zhang, X. Li, W. Ma, Annotating images by mining image search results, IEEE Trans. Pattern Anal. Mach. Intell. (2008) 1919–1932.
[4] H. Jegou, M. Douze, C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 304–317.
[5] Y. Yang, Z. Huang, H. Shen, X. Zhou, Mining multi-tag association for image tagging, World Wide Web 14 (2) (2011) 1–24.
[6] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vision 42 (3) (2001) 145–175.
[7] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide-baseline stereo from maximally stable extremal regions, Image Vision Comput. 22 (10) (2004) 761–767.
[8] E. Tola, V. Lepetit, P. Fua, A fast local descriptor for dense matching, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[9] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 25–32.

[10] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[11] A. Bosch, A. Zisserman, X. Muñoz, Scene classification using a hybrid generative/discriminative approach, IEEE Trans. Pattern Anal. Mach. Intell. (2008) 712–727.
[12] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 524–531.
[13] T. Tuytelaars, Dense interest points, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2281–2288.
[14] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[15] X. Wu, Y. Jia, W. Liang, Incremental discriminant-analysis of canonical correlations for action recognition, Pattern Recognition 43 (12) (2010) 4190–4197.
[16] J. Kim, K. Grauman, Asymmetric region-to-image matching for comparing images with generic object categories, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2344–2351.
[17] C. Gu, J. Lim, P. Arbeláez, J. Malik, Recognition using regions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1030–1037.
[18] S. Todorovic, N. Ahuja, Extracting subimages of an unknown category from a set of images, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 927–934.
[19] T. Malisiewicz, A. Efros, Recognition by association via learning per-exemplar distances, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[20] H. Wang, P. Teng, W. Liang, Packed dense interest points for scene image retrieval, in: IEEE 2011 Sixth International Conference on Image and Graphics (ICIG), 2011, pp. 789–794.
[21] M. Douze, A. Ramisa, C. Schmid, Combining attributes and fisher vectors for efficient image retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2011, p. 6.
[22] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009) 733–746.
[23] J. Tang, S. Yan, R. Hong, G. Qi, T. Chua, Inferring semantic concepts from community-contributed images and noisy tags, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM, 2009, pp. 223–232.
[24] V. Ferrari, A. Zisserman, Learning visual attributes, in: Advances in Neural Information Processing Systems 2008.
[25] X. Zhou, T. Huang, Relevance feedback in image retrieval: a comprehensive review, Multimedia Syst. 8 (6) (2003) 536–544.
[26] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, 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, ACM, 2009, p. 48.
[27] M. Muja, D. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration, in: IEEE Conference on Computer Vision Theory and Applications, vol. 340, 2009, p. 331.
[28] D. Parikh, K. Grauman, Relative attributes, in: ICCV, 2011, pp. 503–510.
[29] D. Dai, T. Wut, S. Zhu, Discovering scene categories by information projection and cluster sampling, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3455–3462.
[30] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Gool, A comparison of affine region detectors, Int. J. Comput. Vision 65 (1) (2005) 43–72.

Han Wang received the B.A. degree in computer science from National University of Defence Technology. She is currently a doctoral candidate in the School of Computer Science at the Beijing Institute of Technology. Her thesis is a focus on the problem of image retrieval and feature extraction. Her research interests include machine learning, computer vision, and image retrieval.

Wei Liang received the B.A. degree in computer science from the Shandong Institute of Light Industry in 2000 and the Ph.D. degree in computer science from the Beijing Institute of Technology. Her research interests include computer vision, human tracking and action perception.

Xinxiao Wu received the B.A. degree in computer science from the Nanjing University of Information Science and Technology in 2005 and the Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. She is currently a lecturer in the School of Computer Science at the Beijing Institute of Technology. Her research interests include machine learning, computer vision, and human action perception.

Peng Teng received the B.A. degree in School of Computer Science at the Beijing Institute of Technology. He is currently a doctoral candidate in the School of Computer Science at the Beijing Institute of Technology. His research interest include machine learning, multi-modal human–computer interaction and signal processing.