Che *et al. EURASIP Journal on Advances in Signal Processing* (2017) 2017:21
DOI 10.1186/s13634-017-0456-1

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Image retrieval by information fusion based on scalable vocabulary tree and robust Hausdorff distance

Chang Che[1,2], Xiaoyang Yu[1], Xiaoming Sun[1*] and Boyang Yu[1]

Abstract

In recent years, Scalable Vocabulary Tree (SVT) has been shown to be effective in image retrieval. However, for general images where the foreground is the object to be recognized while the background is cluttered, the performance of the current SVT framework is restricted. In this paper, a new image retrieval framework that incorporates a robust distance metric and information fusion is proposed, which improves the retrieval performance relative to the baseline SVT approach. First, the visual words that represent the background are diminished by using a robust Hausdorff distance between different images. Second, image matching results based on three image signature representations are fused, which enhances the retrieval precision. We conducted intensive experiments on small-scale to large-scale image datasets: Corel-9, Corel-48, and PKU-198, where the proposed Hausdorff metric and information fusion outperforms the state-of-the-art methods by about 13, 15, and 15%, respectively.

Keywords: Image retrieval, Hausdorff distance, Information fusion, Scalable vocabulary tree

1 Introduction

Image retrieval is an important task in computer vision, which is particularly useful in applications such as Internet image identification for image classification, search and annotation. In recent years, a number of Internet image search systems have been developed [1, 2], which focused on learning a statistical model for mapping image content features to classification labels.

In recent years, a number of deep learning frameworks have been proposed for content-based image retrieval (CBIR). Back-propagation, since 1980s, has been a well-known algorithm for learning the weights of neural networks [3, 4], and widely used in deep learning networks. A typical deep learning approach consists of three phases: (i) training a deep learning model from training data with pre-defined labels; (ii) pass the images through the trained model to extract the feature representations; and finally (iii) applying fully connection layers of the deep architecture or other models such as K-nearest neighbor (KNN) to obtain the best match images. Specifically, for the first stage, several deep architecture of Convolutional Neural Networks (CNNs) can be applied [5–8]. Deep learning approaches for image retrieval achieves the best performance in recent years. However, for a large collection of dataset images, the deep architecture can only be efficiently trained using powerful graphics processing units (GPUs). In contrast, the scalable vocabulary tree (SVT) framework is proven to be a computationally efficient framework [9] for large-scale image retrieval tasks using normal CPUs.

SVT approaches can be regarded as an extension of Bag-of-Words (BoW) approaches since the visual words can be easily extended to tens of thousand at a logarithmic scale [10–13]. In a typical image retrieval framework [1], a scalable vocabulary tree (SVT) is generated by hierarchically clustering local descriptors. First, an initial k-means clustering is performed on the local descriptors of the training data to obtain the first-level k-clusters. Then, the same clustering process is applied recursively to the local descriptors surrounding each cluster at the current level. Repeating this procedure will lead to a hierarchical k-means clustering structure, also known as a SVT in this context. Based on the SVT, the high-dimensional

*Correspondence: xiaoming_66881982@163.com
[1]The Higher Educational Key Laboratory for Measuring and Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, 150080 Harbin, China
Full list of author information is available at the end of the article

histograms of image local descriptors can be generated, which enables efficient image retrieval. When performing image matching on a SVT, local descriptors of the query image are quantized by traversing each layer of the SVT and a histogram over the tree nodes (visual words) is generated. Candidate images are then sorted according to the similarity of these histograms to the query image histogram. A histogram of the local descriptors is called image signature in this context.

Although SVT-based approaches normally produces good results, there are still potential for us to further improve the performance. For example, Hausdorff distance [14–16] is a popular post-processing approach which is used to match image signatures between the query image and database images. Using the conventional Hausdorff distance metric [14] as the image feature matching technique, the maximum local distance is chosen as the matching distance, which produces low distance even if every pair of image feature elements are close between a query image and a dataset image. However, it is sensitive to noise, i.e., the image local descriptor from the cluttered background. The limited improvement achieved when compared to its extra computational cost make it undesirable in large-scale image retrieval. In view of this, we propose to utilize an improved Hausdorff distance for the refinement of SVT matching results. In addition to the new distance metric, a new framework is proposed for large-scale image retrieval which can fuse matching results from different image representation methods. Each individual image representation adopted in the framework has the flexibility to utilize any one of the different image signatures based on SVT.

Specifically, two new algorithms are developed in the framework: (1) an improved Hausdorff Distance algorithm which helps remove outliers and improve retrieval accuracy in the proposed framework; and (2) a fusion algorithm that can combine the matching results generated from different image representation methods, and produces the final matching list. To the best of our knowledge, this is the first work to utilize Hausdorff distance into the SVT-based image retrieval systems.

Experimental results show that, by embedding and fusing different image representation methods in the proposed framework, the image retrieval performance is superior to using each image representation method alone. Experimental results show that, relative to a baseline approach using SVT built upon SIFT descriptors, utilizing Hausdorff distance improves more than 10% of retrieval performance with negligible computational cost.

2 Proposed image retrieval framework

We proposed a new image retrieval framework that incorporate various sources of information relative to the conventional image retrieval based on an individual image representation, as shown in Fig. 1. Based on various local descriptors and visual word vocabularies, various image representations produce various image search lists from the database images. Using the information fusion algorithm, different sets of searching results can be fused to produce the final matching list. The final image retrieval performance is expected to be superior to that produced by each search method since it fuses different sources of information.

In order to evaluate the effectiveness of the proposed fusion algorithm, several image representations are performed here, as shown in Fig. 2. Specifically, three image search methods are used: (i) SVT approach based on histogram of dense-patch SIFT descriptors; (ii) SVT approach based on kernel density of dense-patch SIFT descriptors; and (iii) SVT approach based on histogram of dense-patch DAISY descriptors. The final image retrieval performance will be validated to be superior to that produced by any of the individual matched image list.

For an individual image representation, we propose a new image retrieval framework that incorporate the kernel density information and a robust Hausdorff distance metric, as shown in Fig. 3. In the offline training phase, kernel density, which is optional, is incorporated in two stages: vocabulary tree construction and image histogram representation; in the online recognition phase for the query image, the kernel density is only involved in the histogram calculation. To keep the efficiency of the retrieval system, the offline training phase is kept the same as the SVT image retrieval; while in the online retrieval phase for the query image, the Hausdorff distance is only involved in the refinement of SVT image retrieval.

3 Proposed algorithms

3.1 Feature extraction

Amongst the local features, scale-invariant feature transform (SIFT) is most widely used in recent years, and some variants of SIFT have also been proposed. The keypoint-based SIFT descriptor developed in [17, 18] is one of the most popular local features [19–21]. According to the comparative study in [22], SIFT is generally superior to other descriptors, such as moment invariants [23] and shape context [24], due to its robustness to affine transformations and illumination changes.

During SIFT descriptor extraction, the difference of Gaussian (DoG) space of the image is calculated first. Then, the keypoints are detected by finding those invariant to different scales in the DoG space. Based on the local region surrounding the detected keypoints, the orientation histogram is computed as the local descriptor. SIFT descriptors are relatively robust to image noises, illumination changes, as well as limited changes in viewing angles of the object.

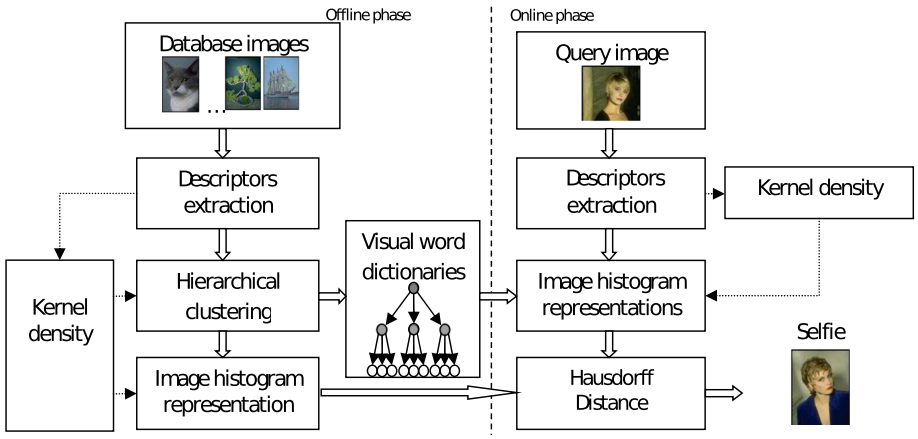

Fig. 1 Overview of the proposed image retrieval framework

To reduce computational cost, SIFT descriptors can also be extracted based on regular patches rather than the detected interest keypoints, which is usually called dense patch SIFT or dense sampling SIFT. Since the detected SIFT keypoints are robust to scales and rotations, the sparse SIFT descriptors are commonly used in general object matching. However, it has been reported that dense SIFT outperform sparse SIFT in some applications [25–30]. According to the survey made in [31], SIFT descriptors based on dense overlapping regular image patches is promising among state-of-the-art image feature extraction methods. According to the preliminary experiments in this work, sparse SIFT descriptors will produce about 16% less precision than that of dense-patch SIFT descriptors, which is consistent to the experimental results in [31]. Therefore, we only extract dense-patch descriptors for the image retrieval task.

DAISY is another state-of-the-art image descriptor [32, 33] which needs about only one tenth of the number of computational operations of SIFT descriptors [33]. DAISY is essentially similar to SIFT, except that is uses a Gaussian kernel to aggregate the gradient histograms in

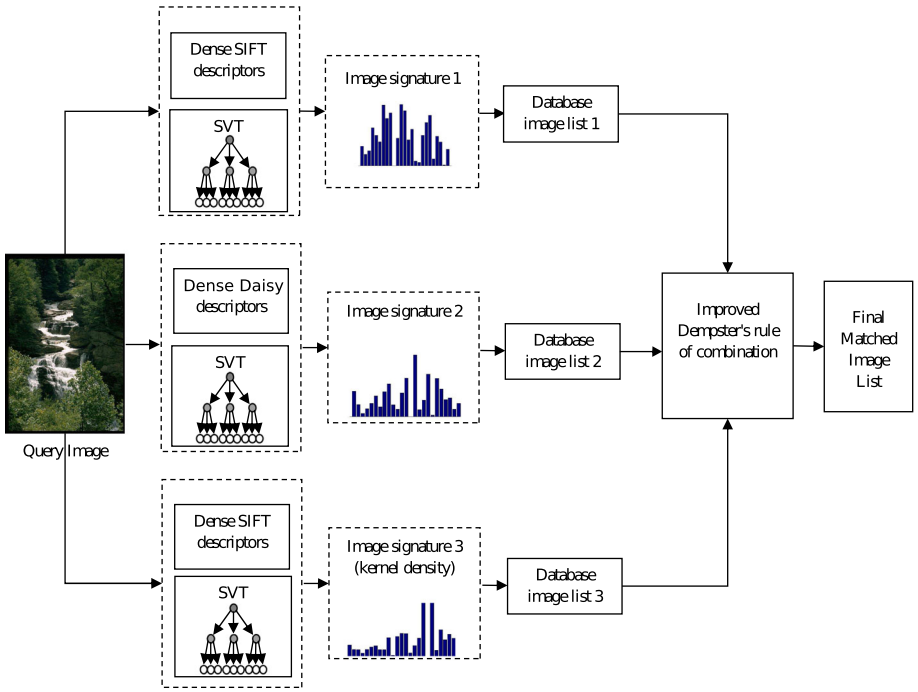

Fig. 2 Proposed fusion framework embedded with the proposed image signatures

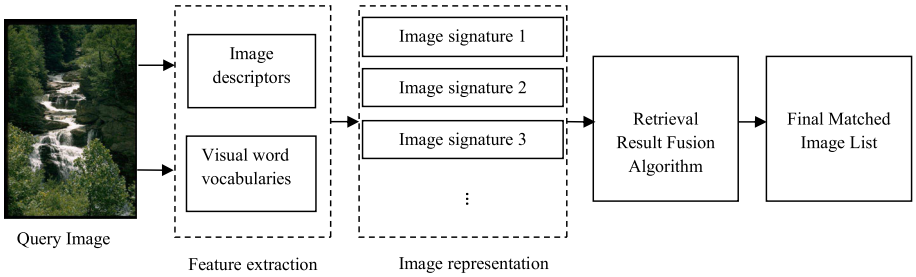

Fig. 3 Proposed framework for image retrieval based on individual image representations

different bins whereas SIFT relies on a triangular shaped kernel. The performance of the dense patch daisy descriptors is comparable to dense patch sift descriptors. However, SIFT is found to be still outperforming DAISY in some application domains [34].

3.2 Image representation

Towards large-scale image retrieval, the Scalable Vocabulary Tree (SVT) model [1] is well exploited in the state-of-the-art works [1, 2, 9, 28, 35]. A SVT T is constructed by hierarchically clustering of local descriptors, which consists of $C = B^L$ codewords, where B is the branch factor and L is the depth. Let each node $\mathbf{v}^{l,h} \in T$ in the tree represents a visual codeword, where $l \in \{0, 1, \cdots, L\}$ indicates its level, and $h \in \left\{1, 2, \cdots, B^{L-l}\right\}$ is the index of its node in its level. A query image q is represented by a bag of N_q local descriptors $\mathbf{X}_q = \left\{\mathbf{x}_{q,i}\right\}, i \in N_q$. Each $\mathbf{x}_{q,i}$ is traversed in T from the root to a leaf to find the nearest node, resulting in the quantization $T(\mathbf{x}_i) = \left\{\mathbf{v}_i^{l,h}\right\}_{l=1}^{L}$. Thus, a query image is eventually represented by a high-dimensional sparse Bag-of-Words (BoW) histogram $\mathbf{H}_q = [h_1^q, \ldots, h_C^q]$ obtained by counting the occurrence of its descriptors on each node of T.

After clustering the SIFT descriptors of training image patches by SVT, we obtain the codewords which are the cluster centers. For each codeword $\mathbf{c}$ in the codebook $\mathcal{CB}$, traditional codebook model estimates the distribution of codewords in an image by a histogram as follows:

$$\mathrm{H}(\mathbf{c}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(\mathbf{c} = \arg\min_{\mathbf{w}} \left(D(\mathbf{w}, \mathbf{v}_i)\right)\right)_{\mathbf{w}\in\mathcal{CB}} \quad (1)$$

where N is the number of patches in an image, $\mathbf{v}_i$ is the descriptor of an image patch, $D(\cdot,\cdot)$ is the Euclidean distance, and $\mathrm{I}(\cdot)$ is the identity function.

A robust alternative to histograms for estimating a probability density function is kernel density estimation (KDE) by [36]. KDE uses a kernel function to smooth the local neighborhood of data samples. KDE is advantageous over histogram. First, its nonparametric nature provides us with enough flexibility to model feature distributions for a broad and diverse set of scenes. Second, in contrast to the histogram estimator, its smoothing parameter can be adjusted to make the descriptors relatively insensitive to small descriptor variations and to imperfections in scale normalization. Third, descriptors can still be computed very efficiently when KDE is coupled with the fast Gauss transform (FGT) by [37]. A high-dimensional estimator with kernel K and bandwidth parameter $\mathbf{B}$ is given by

$$f(\mathbf{c}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{K}_{\mathbf{B}}(\mathbf{v}_i - \mathbf{c}) \quad (2)$$

In this paper, we use the SIFT descriptor that draws on the Euclidean distance as its distance function. Since the Euclidean distance that we measure between SIFT descriptors assumes a Gaussian distribution, Gaussian-shaped kernel is adopted here:

$$\mathrm{K}_{\mathbf{B}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{B}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{B}^{-1}\mathbf{x}\right) \quad (3)$$

where m is the dimensionality of the descriptor, and the bandwidth parameter matrix $\mathbf{B} \in \mathcal{R}^{m\times m}$ models the degree of uncertainty about the sources and controls the smoothing behavior of the KDE.

We use 10-fold cross-validation for determining the optimal parameters. Hence, the size of the kernel is dependent on the dataset and the image descriptor. In order to reduce the computational cost, we simplify by making all the diagonal elements in $\mathbf{B}$ to have the same value and all off-diagonal elements to be zeros. We split the training set into 10 roughly equal sized parts. For each setting of parameters and, using nine parts we fit the parameter, and calculate the retrieval precision on the remaining one part as the validation set. We repeat this procedure by using every part as the validation set in each of the 10 runs. Finally we get an average of the 10 precisions which corresponds to the setting. We choose the setting of which corresponds to the maximum average precision.

We first construct SVT by hierarchically clustering of local descriptors, i.e., SIFT descriptors. The SVT T consists of $C = B^L$ codewords, where B is the branch factor and L is the depth. Then, a query image q is represented by a bag of N_q local descriptors $\mathbf{X}_q =$

$\{\mathbf{x}_{q,i}\}, i \in N_q$. Each $\mathbf{x}_{q,i}$ is traversed in T in all its leaves to find the kernel density $f(\mathbf{c})$ for each leaf, resulting in the kernel density vector $F(\mathbf{x}_i) = \{f(\mathbf{c}_j)\}_{j=1}^{B^L}$. Thus, a query image is eventually represented by a high-dimensional real-valued Bag-of-Words (BoW) histogram $\mathbf{F}_q = [f_q(\mathbf{c}_0), f_q(\mathbf{c}_1), \ldots, f_q(\mathbf{c}_{B^L})]$ obtained by calculating kernel density on each leaf node of T. The kernel density descriptors of the database images are denoted by $\{\mathbf{F}_{d_m}\}_{m=1}^{M}$, where M is the total number of database images.

Based on dense patch DAISY and dense patch SIFT, we have two sets of image signatures (representations): $\left\{\mathbf{F}_q, \{\mathbf{F}_{d_m}\}_{m=1}^{M}\right\}$ and $\left\{\mathbf{G}_q, \{\mathbf{G}_{d_m}\}_{m=1}^{M}\right\}$, respectively. In addition, the basic image signature (BoW histogram) based on dense patch SIFT is denoted as $\left\{\mathbf{H}_q, \{\mathbf{H}_{d_m}\}_{m=1}^{M}\right\}$. Image signatures $\mathbf{H}_q$, $\mathbf{F}_q$, and $\mathbf{G}_q$ are B^L-dimensional vectors, and typically B^L ranges from 10^4 to 10^6, as suggested in [9]. We will show how the three types of image representations can be fused to achieve better image retrieval performance than any single image representation.

3.3 An improved Hausdorff metric for image matching

Denote X and Y for two sets of vectors: $X = \{x_i\}, i = 1, 2, \cdots, M$, and $Y = \{y_i\}, i = 1, 2, \cdots, N$, where x_i and y_i are both D-dimensional vectors. Then, the Hausdorff distance can be defined as a root mean square distance as follows:

$$d_{\text{root}}(X, Y) = \sqrt{\frac{1}{|X|}\int_{x\in X} d(x, Y)^2 \, dX} \tag{4}$$

Conventionally, in order to reduce the computational complexity, the Hausdorff distance $d_0(X, Y)$ is defined by

$$d_0(X, Y) = sup\{\sup_{x\in X}\inf_{y\in Y} d(x, y), \sup_{y\in Y}\inf_{x\in X} d(x, y)\} \tag{5}$$

where *sup* represents the supremum and *inf* the infimum.

After the image representation, each image $\mathbf{I}$ can be represented by a B^L-dimensional signature in real-valued domain R^{B^L}. Denote the signature of the query image and a database image be X and Y, respectively. Then, the Hausdorff distance between X and Y can be regarded as the measure for the image retrieval. First, we define the distance $d(x, Y)$ between a point x belonging to the set X and the set Y as:

$$d(x, Y) = min_{x\in X} \|x - y\|_2 \tag{6}$$

and

$$d(X, Y) = \max\{max_{x\in X} min_{y\in Y} d(x, y), max_{y\in Y} min_{x\in X} d(x, y)\} \tag{7}$$

Informally, two sets are close in the Hausdorff distance if every point of either set is close to some point of the other set. The Hausdorff distance is the greatest of all the distances from a point in one set to the closest point in the other set. However, this basic metric is not robust since a few outliers will affect the *max* operation result. For example, if only one point $y \in Y$ is far away from all other points which are all similar in X and Y, then $d(X, Y)$ will be large. This is likely to happen when a few visual words of SVT origin from the cluttered background, and the local SIFT descriptors of the query image may have high occurrences over these visual words, resulting in high Hausdorff distances to the database images which have the same foreground objects but different backgrounds.

In order to diminish the effect of the outliers, the directed distance $d_{\text{H}}(X, Y)$ of the proposed Hausdorff distance is proposed by replacing the Euclidean distance by the cost function:

$$d_{\text{H}}(X, Y) = \frac{1}{|X|}\sum_{x\in X} \gamma(d(x, Y)) \tag{8}$$

where $|X|$ is the cardinality of the image signature X, and the cost function $\gamma(t)$ is convex and symmetric and has a unique minimum value at zero. In our experiments, we use the cost function defined by

$$\frac{d\gamma(t)}{dt} = k \cdot \gamma(t)\left(1 - \frac{\gamma(t)}{\tau}\right), \gamma(0) = \gamma_0 \tag{9}$$

By the above definition, when the distance $d(x, Y)$ is small, $\gamma(d(x, Y))$ is very small, diminishing the effect of random noises; when the distance $d(x, Y)$ is larger, $\gamma(d(x, Y))$ becomes large, reflecting the actual distances between the points; when the distance $d(x, Y)$ is very large, $\gamma(d(x, Y))$ gets limited by a threshold, controlling the effect of the outliers that may dramatically increases the distance.

By solving the differential Eq. (9), we get

$$\gamma(t) = \frac{\tau}{1 + \left(\frac{\tau}{\gamma_0} - 1\right) exp(-kt)} \tag{10}$$

where $0 < \gamma_0 < 1$ is set to be 0.1 experimentally, $k = 0.05$, and $\tau = 0.8$ is a threshold to eliminate outliers, so the outliers yielding large distances are diminished. Since the matching performance depends on the parameter τ, it is important to determine it appropriately. If it is set to infinity, this proposed Hausdorff distance is equivalent to the conventional one. Because the cost function is associated with the distance value $d(x, Y)$, the threshold value is selected experimentally. The function $\gamma(t)$ is illustrated in Fig. 4.

3.4 Image match scoring

For the purpose of image retrieval, image descriptors need to be indexed by similarity scoring. The database images are denoted by $\{d^m\}_{m=1}^{M}$, where d^m is a local descriptor, and M is the number of images in the database. Following

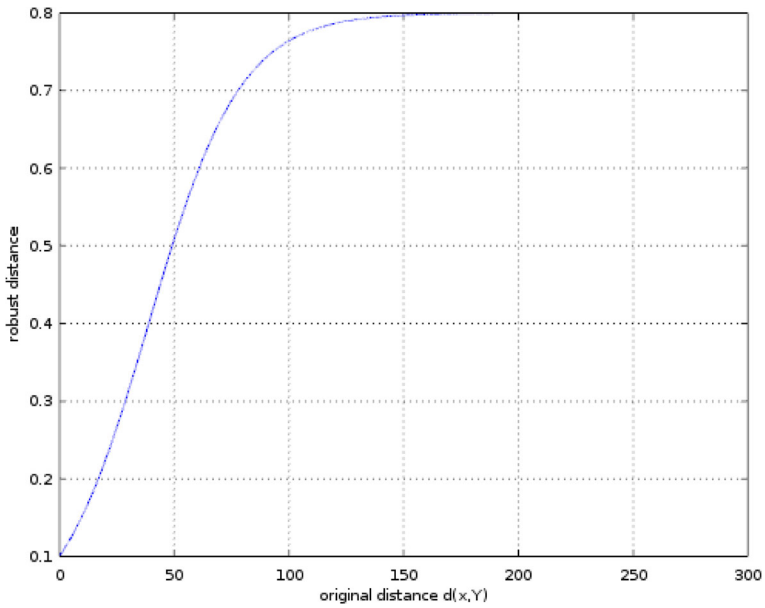

Fig. 4 Components of the calculation of the Hausdorff distance between the *green line X* and the *blue line Y*

the same VT quantization procedure, the local descriptors $\{y_j\}$ in d^m are mapped to a high-dimensional sparse Bag-of-Words (BoW) histogram $\mathbf{H}_d = [h_1^d, \ldots, h_C^d]$. The images with highest similarity score $sim(q, d^m)$ between query q and database image d^m are returned as the retrieval result. Conventionally, the similarity $sim(q, d^m)$ is defined in [1] as

$$sim\left(\mathbf{H}_q, \mathbf{H}_{d^m}\right) = 1 - \left\| \frac{\mathbf{H_q} \cdot \mathbf{w}}{\left\|\mathbf{H}_q \cdot \mathbf{w}\right\|} - \frac{\mathbf{H}_{d^m} \cdot \mathbf{w}}{\|\mathbf{H}_{d^m} \cdot \mathbf{w}\|} \right\| \quad (11)$$

where $\mathbf{w} = [w_1, w_2, \cdots, w_C]$, $w_i = \ln \frac{M}{M_i}$, and M_i is the number of images in the database with at least one descriptor vector path through node i.

Since the vocabulary tree is very large, the number of images whose descriptors are quantized to a particular node is normally zero. In [1], the scalability is addressed by only comparing those database images indexed by each non-zero codeword for the given query image. Here, $sim\left(\mathbf{H}_q, \mathbf{H}_{d^m}\right)$, $m = 1, 2, \ldots, M$ is normalized with a *Sigmoid* function to enhance the larger similarity scores:

$$p_m = \frac{1}{1 + e^{-\alpha sim_m}} \quad (12)$$

where α is the scaling parameter that is set to be 10, which produces the best performance according to preliminary experiments. Here, we use 10-fold cross-validation for determining the optimal parameters. We split the training set into 10 roughly equal sized parts. The setting of which corresponds to the maximum average precision is chosen.

3.5 Proposed information fusion

Let the node ensemble contain SVT nodes $T_i, i = 1, 2, \ldots, N$, where N is the total number of SVT nodes, e.g., $N = B^L$. By applying SVT approach to the query image, there are M similarity scores for the query image. An example for tag fusion of the three density score lists is illustrated in Fig. 5, where s_1 , s_2 and s_3 are the three similarity score lists generated by three types of image representation, s is the final score list to be fused. After obtaining the final score list s, it is sorted in descending order and the top m, $m < M$ nodes are suggested for the query image.

To integrate the three lists of scores, Dempster's rule of combination [38] is utilized to combine different sources because it is considered to be a more flexible and general approach than the traditional probability theory and it is able to deal with some ignorance of the system. The basic probability assignment (BPA) function is used here to take into account all the available evidence, and is defined as a mapping S from the power set 2^Ω of a finite set $\Omega = \{A_1, A_2, \cdots, A_N\}$ to $[0, 1]$ that for any $T \in 2^\Omega$, and we have

$$\sum_{T \in 2^\Omega} S(T) = 1,\ S(T) \geq 0 \quad (13)$$

where the power set 2^Ω comprises of exhaustive set of mutually exclusive elements:

$$2^\Omega = \{\{A_1\}, \cdots, \{A_{N_t}\}, \{A_1, A_2\}, \cdots, \{A_{N_t-1}, A_{N_t}\}, \cdots, \{A_1, \cdots, A_N\}, \phi\} \quad (14)$$

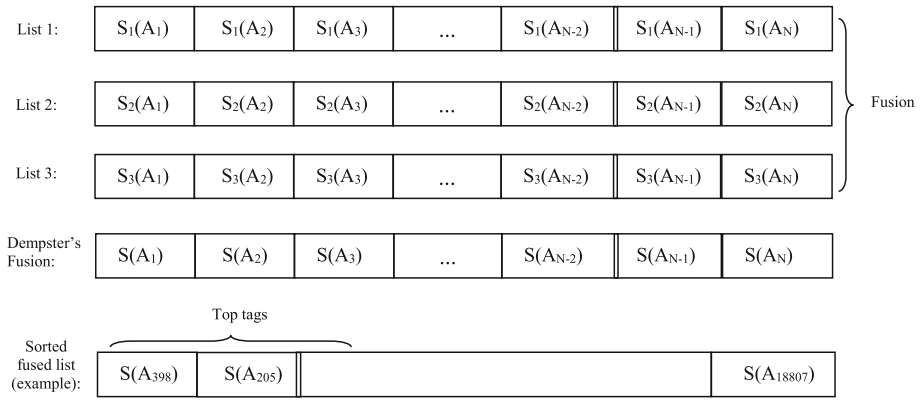

Fig. 5 Overview of the tag fusion process

where ϕ is the empty set, $S(\phi) = 0$ in a close-world assumption, and there are in total 2^N elements in 2^Ω.

Dempster's rule for combining K sources is:

$$S(T) = \frac{\sum_{T_1,T_2,\cdots,T_K \subset 2^\Omega, \cap_{i=1}^K T_i = T} (S_1(T_1) \cdots S_K(T_K))}{\sum_{T_1,T_2,\cdots,T_K \subset 2^\Omega, \cap_{i=1}^K T \neq \phi} (S_1(T_1) \cdots S_K(T_K))} \tag{15}$$

In this work, the joint BPA for three sources is reformulated as follows:

$$S(T) = \sum_{T_1,T_2,T_3 \subset 2^\Omega, T_1 \cap T_2 \cap T_3 = T} \frac{S_1(T_1) S_2(T_2) S_3(T_3)}{1 - M} \tag{16}$$

where $M = \sum_{T_1 \cap T_2 \cap T_3 = \phi} (S_1(T_1) S_2(T_2) S_3(T_3))$ is a measure of the amount of conflict among the three BPA sets.

In order to satisfy the fast-response-time requirement of the image retrieval, instead of directly using Dempster's rule of combination, we improve the online procedure. Originally, the BPAs need to be estimated on as many as 2^N elements in the power set 2^Ω, and the computational complexity is as high as $O(2^N)$ which is not affordable in real applications. Here, we reduce the original power set 2^Ω to a much smaller subset

$$P = \{\{A_1\}, \cdots, \{A_{N_t}\}, \{A_1, A_2\}, \cdots, \{A_{N-1}, A_N\}, \Phi\} \tag{17}$$

where Φ is a subset of 2^Ω which contains the elements with more than two SVT nodes. The reduced set P

Fig. 6 Sample images from the Oxford Building-11 dataset

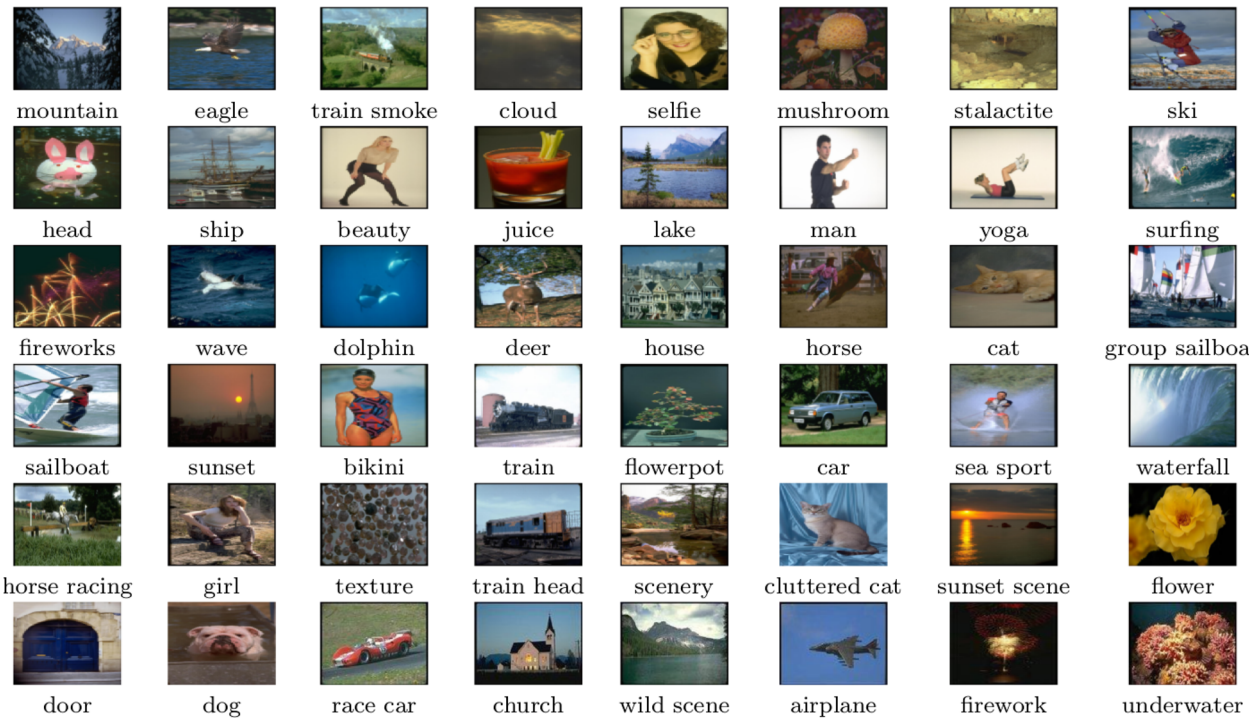

Fig. 7 Sample images from the Corel-48 dataset

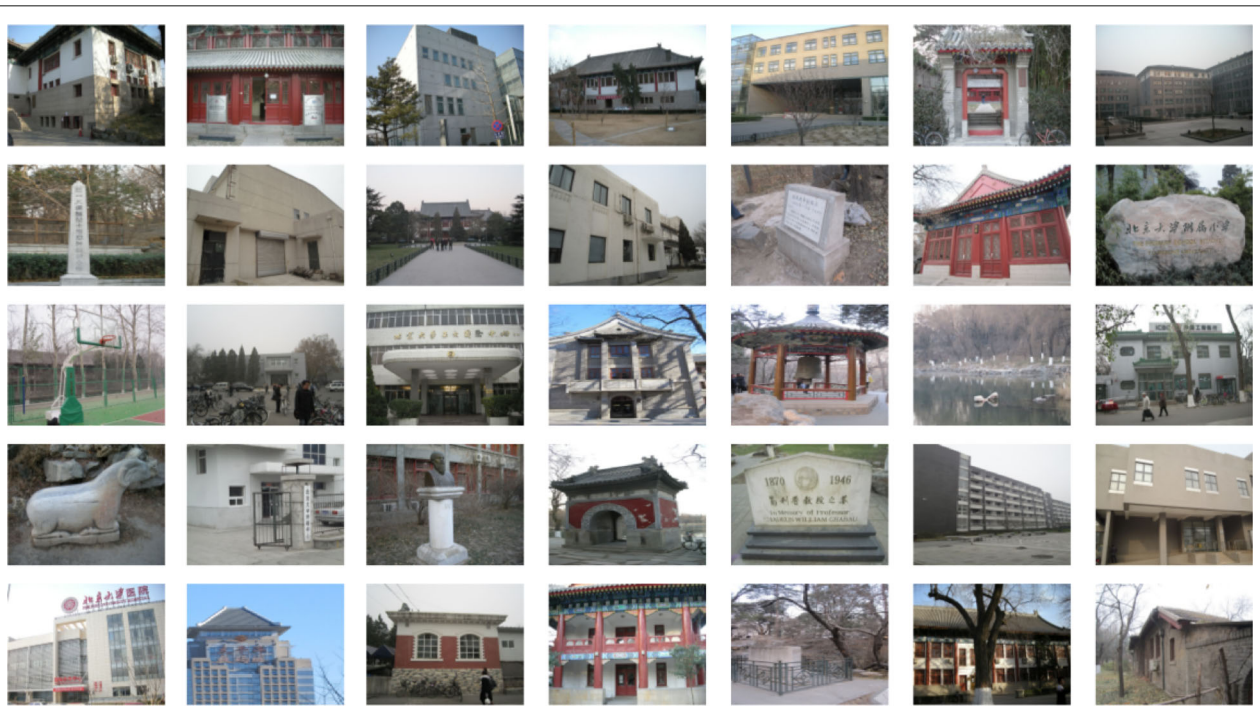

Fig. 8 Sample images of PKU landmark-198 dataset

has $N^2 + 1$ elements so the computational complexity is $O\left(N^2\right)$. The computational cost will be reduced substantially since $N = B^L$ here. So, the joint BPA can be simplified as follow:

$$S(T) = \sum_{T_1,T_2,T_3 \subset P, T_1 \cap T_2 \cap T_3 = T} \frac{S_1(T_1) S_2(T_2) S_3(T_3)}{1 - M} \tag{18}$$

where $M = \sum_{T_1 \cap T_2 \cap T_3 = \phi} (S_1(T_1) S_2(T_2) S_3(T_3))$ and the BPAs on each element of P for each source can be formulated as follows:

$$S_t(A_i) = S_t(\{A_i, A_i, A_i\}) = m_t(\{A_i\})^3 \tag{19}$$

$$S_t\left(\left\{A_i, A_j, A_k\right\}\right) = m_t(\{A_i\}) m_t\left(\left\{A_j\right\}\right) m_t(\{A_k\}) \tag{20}$$

where $A_i \subset \Omega$, $t = 1, 2, 3$ and $S(\Phi)$ denote the ignorance on the power set, is set to be 0.15 empirically.

For any $T_t \subset P$, $k = 1, 2$, we can estimate $S_1(T_t), S_2(T_t)$ and $S_3(T_t)$ according to (19) and (20), and then we calculate $S(A_i)$ for every $A_i \subset \Omega$ according to (18). Finally, we sort $S(A_i)$, $A_i \subset \Omega$ descendingly and get the ranked top database image as the search result.

3.6 Proposed image search algorithm with robust Hausdorff distance

Using the proposed algorithms, a matching image list from the database will be generated from a query image. The algorithm is briefed in Algorithm 1.

Algorithm 1 Proposed Image Search Algorithm with Robust Hausdorff Distance

Input: A set of N SIFT descriptors $\mathbf{S} = \{\mathbf{s}_1, \cdots, \mathbf{s}_N\}$ along with their kernel densities $\mathbf{w} = \{w_1, \cdots, w_N\}$; **Output**: Final matching score vector for the m-th image over the K image categories is $\mathbf{s}_m = \{s_1, \cdots, s_N\}$; The top matched image in the database is indexed by the largest value of s_m: $c_m^* = \arg\max_k S_{m,k}^{(L)}$. **Initialization**: The vocabulary tree T consisting of $C = B^L$ visual codewords $\mathbf{v}^{l,h} \in T$, where $l \in \{0, 1, \cdots, L\}$ indicates its level, and $h \in \left\{1, B^1, \cdots, B^{L-l}\right\}$ is the its index at its level. **Repeat**: For each of the three image representation method:

1. Generate image representation according to 3.2;
2. Use the proposed Hausdorff metric to perform image matching according to 3.3;
3. Obtain the normalized image matching score vector according to 3.4;

Apply the proposed information fusion algorithm in 3.5 to obtain the final image matching list in database.

Table 1 Comparison of various SVT approaches on Oxford Building-11 dataset

Scheme	Signature			Top 1	Top 3	Top 5
	Baseline SVT	Density	Hausdorff			
(a)	✓	×	×	0.697	0.776	0.788
(b)	✓	✓	×	0.713	0.798	0.802
(c)	✓	×	✓	0.733	0.807	0.821
(d)	✓	✓	✓	0.747	0.821	0.834

4 Experimental results

4.1 Datasets

4.1.1 Oxford Building-11 Dataset

The Oxford Building Dataset [39] comprises of 5062 images collected from Flickr by searching for particular Oxford landmarks such as "All Souls Oxford" and "Christ Church Oxford", as shown in Fig. 6. The collection is manually categorized into 11 different landmarks, and the query set contains 55 images. This is a challenging benchmark for object search due to occlusion and cluttered background.

4.1.2 Corel natural image dataset

The natural landscape images used in this work include a total of 4798 images, as shown in Fig. 7, derived from Corel photo CDs. The dataset images are randomly partitioned in 7:3 ratio, for training set and test set, respectively.

4.1.3 PKU landmark dataset

We also test our proposed algorithms on the landmark benchmark from MPEG CDVS requirement subgroup [35] , which contains 13,179 scene photos, organized into 198 landmark locations from the Peking University Campus. Sample images of PKU landmark-198 dataset are given in Fig. 8. The dataset images are randomly partitioned in two halves for training set and test set.

4.2 Experiment settings

For efficient dense patch SIFT descriptor extraction, we sampled on overlapping 16×16 pixel patches in space of 8 pixels [27] for all the algorithms on all the datasets. We adopt the standard DAISY setting as radius $R = 15$,

Table 2 Comparison of various SVT approaches on Corel-48 dataset

Scheme	Signature			Top 1	Top 3	Top 5
	Baseline SVT	Density	Hausdorff			
(a)	✓	×	×	0.603	0.680	0.698
(b)	✓	✓	×	0.627	0.702	0.721
(c)	✓	×	✓	0.640	0.718	0.735
(d)	✓	✓	✓	0.662	0.739	0.752

Table 3 Comparison of various SVT approaches on PKU dataset

Scheme	Signature			Top 1	Top 3	Top 5
	Baseline SVT	Density	Hausdorff			
(a)	✓	×	×	0.551	0.634	0.652
(b)	✓	✓	×	0.579	0.649	0.671
(c)	✓	×	✓	0.595	0.670	0.691
(d)	✓	✓	✓	0.616	0.698	0.715

radius quantization levels $Q = 3$, angular quantization levels $T = 8$, and histogram quantization levels $H = 8$ as in [33].

All images are resized to 640 × 480 resolution as a tradeoff between image retrieval efficiency and accuracy. In SVT-based approaches, the branch number is set to be 10 and vocabulary tree depth is set to be 6. According to the setting in [9], this setting produces satisfactory performance on large-scale image datasets. For matching of SVT histograms between the query images and database images, we adopt the intersection kernel [40] due to its efficiency. The multi-class classification method is the "one-against-all" method [41]. The simulation environments are given as follows: Ubuntu 14.04, Intel® Core™ i7-3770S CPU @ 3.10 GHz x 8, 8-G RAM.

4.3 Preliminary performance evaluation

The proposed image retrieval framework has the flexibility to enhance in several stages: feature extraction, image representation and image matching. We test our proposed approach progressively in such a way: baseline SVT based on dense patch SIFT descriptors, baseline SVT with kernel density to obtain image signature, baseline SVT incorporated with robust Hausdorff distance, and baseline SVT incorporated with both kernel density and robust Hausdorff distance. To evaluate the retrieval performance, we report the retrieval rate in terms of the number of the top returned categories. The query image is regarded to be correctly recognized when its best matched image corresponds to one of the top n returned categories.

The comparison of the above progressive approaches in terms of different top n matched categories are shown in Tables 1, 2, and 3. We can observe that, incorporating the information in either retrieval stages will progressively improve the retrieval performance. It is obvious that integrating baseline SVT approach with kernel density and robust Hausdorff distance is the best choice.

Using the optimal SVT approach above, three image signatures (representations) can be derived: (i) signature (1st) based on histogram of dense-patch SIFT descriptors; (ii) signature (2nd) based on kernel density of dense-patch SIFT descriptors; and (iii) signature (3rd) based on histogram of dense-patch DAISY descriptors. The comparison of the above signatures as well as the fused result using Dempster's rule of information fusion are shown in Tables 4, 5, and 6 in terms of different top n matched categories.

Figure 11 gives the performance comparison for PKU-198 dataset in terms of precision. By using the proposed fusion algorithm, the final retrieval results consistently outperforms the baseline SVT approach by about 15% in terms of retrieval precision.

We can observe that, the 1st signature, i.e., the scheme (c) in Tables 1, 2 and 3, produces moderate performance. The 2nd signature, i.e., the scheme (d), produces better performance due to the incorporation of the kernel density in signature generation stage. The 3rd signature, i.e., the scheme (d) embedded with dense-patch DAISY descriptor instead of dense patch SIFT descriptor, produces slightly inferior performance than SIFT. It is obvious that integrating all the three signatures using the proposed information fusion algorithm is the best choice, which is superior to the retrieval performance of any individual signature.

It is observed that the lowest retrieval performance is produced from PKU-198 dataset since the number of categories is large and geometrically different landmarks may have similar appearance. Oxford-11 produces relatively low performance although the number of categories is

Table 5 Comparison of various image representations on Corel-48 dataset

	Scheme	Top 1	Top 3	Top 5
(a)	1st signature	0.640	0.718	0.735
(b)	2nd signature	0.662	0.739	0.752
(c)	3rd signature	0.642	0.717	0.733
(d)	Fusion result	0.791	0.870	0.895

Table 4 Comparison of various image representations on Oxford Building-11 dataset

	Scheme	Top 1	Top 3	Top 5
(a)	1st signature	0.733	0.807	0.821
(b)	2nd signature	0.747	0.821	0.834
(c)	3rd signature	0.728	0.795	0.819
(d)	Fusion result	0.864	0.928	0.945

Table 6 Comparison of various image representations on PKU-198 dataset

	Scheme	Top 1	Top 3	Top 5
(a)	1st signature	0.595	0.670	0.691
(b)	2nd signature	0.616	0.698	0.715
(c)	3rd signature	0.596	0.669	0.687
(d)	Fusion result	0.743	0.813	0.838

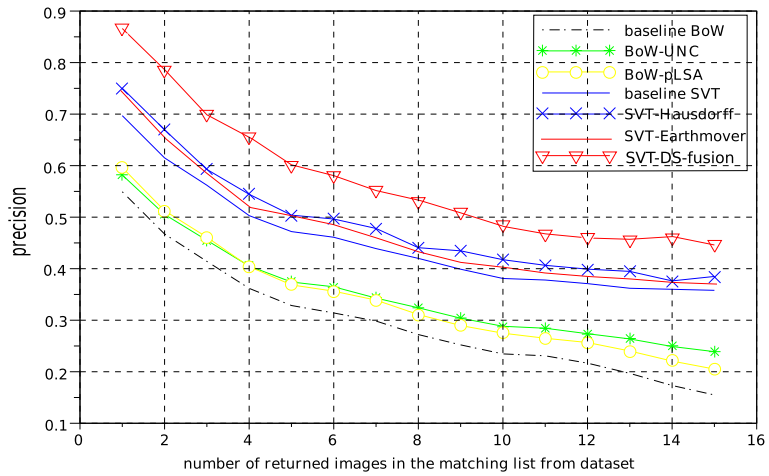

Fig. 9 Oxford-11 performance in terms of precision

small, because of the challenging occlusions and changes in viewpoints and scales. Corel-48 is a medium scale dataset, which produces a moderate performance.

4.4 Objective performance comparisons

To further evaluate the retrieval performance, various methods for image retrieval are tested on the three datasets. The image search performance is evaluated in terms of precision. For each query image, the top matched image list generated by the retrieval system are compared with the ground-truth categories in the dataset. Denote the set of matched images returned by the retrieval system to be S and the set of ground-truth matched images from the database to be T, then precision is defined as

$$Precision = \frac{|S \bigcap T|}{|S|} \tag{21}$$

where $|\cdot|$ denotes the cardinality of the set.

The methods include the baseline BoW approach in [42], the codeword uncertainty (BoW-UNC) method [42], Bosch's hybrid BoW-pLSA method [43], baseline SVT approach [1], SVT-Earthmover (baseline SVT approach combined with Earth Mover's Distance [44] for image matching), the proposed optimized SVT-Hausdorff approach (with kernel density to obtain SIFT signature and Hausdorff distance for image matching), and the proposed SVT-fusion approach with DS fusion (from all the three image matching score lists). Earth mover's distance is a method to evaluate distance between two multi-dimensional distributions by linear programming [44]. UNC uses kernel density estimation to replace the hard-assignment BoW histogram, and reduces the effect of quantization. UNC-SVM and pLSA-SVM are both state-of-the-art methods to fuse generative models and

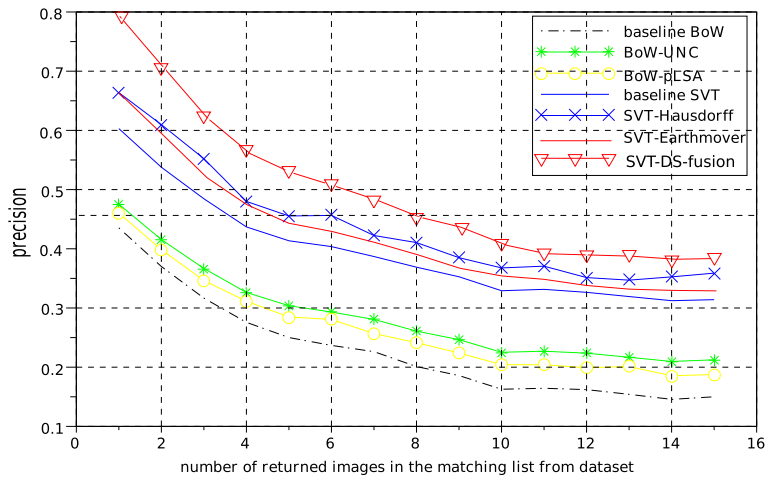

Fig. 10 Corel-48 performance in terms of precision

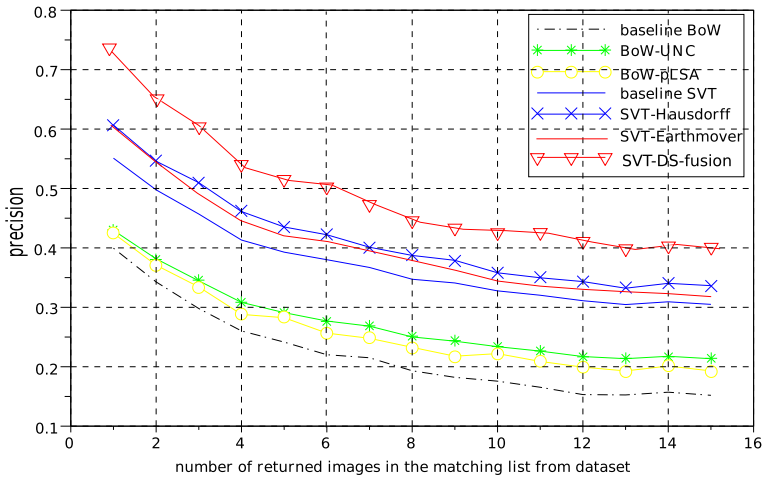

Fig. 11 PKU-198 performance in terms of precision

discriminative models, where the parameters are optimized following the literature to achieve the highest recognition performance.

Figure 9 gives the performance comparison for Oxford-11 dataset in terms of precision. From Fig. 9, it is obvious that the proposed image retrieval approach consistently outperforms other approaches for various number of suggested tags. It is noticed that the UNC-SVM and pLSA-SVM approaches are consistently superior to the baseline BoW-SVM approach, but inferior to the baseline SVT approach. This indicates that SVT-based approaches are generally superior to BoW-based approach. The proposed image retrieval approach consistently outperforms other approaches for various number of suggested tags. By using the proposed fusion algorithm, the final retrieval results consistently outperforms the baseline SVT approach by about 13% in terms of retrieval precision.

Figure 10 gives the performance comparison for Corel-48 dataset in terms of precision. By using the proposed fusion algorithm, the final retrieval results consistently outperforms the baseline SVT approach by about 15% in terms of retrieval precision.

From Figs. 9, 10, and 11, it is observed that the proposed modified Hausdorff distance can significantly improves baseline SVT approach by about 5%, and outperforms the conventional Earth Mover's distance by about 1%.

5 Conclusions

This paper presents a new framework for image retrieval, which has the sufficient flexibility to incorporate various enhancements based on kernel density, robust Hausdorff distance, and information fusion. They are carried out in the stages of image signature generation, image matching, and final scoring list, respectively. By embedding various signatures in the proposed framework, the final retrieval performance is superior to each individual approach. Experimental results show that the proposed framework significantly outperform state-of-the-art content-based image retrieval approaches. Future work may include the integrating other types of context information to the content analysis.

Acknowledgements
This study was supported by the National Natural Science Foundation of China (61401126), National Natural Science Foundation of China (F011702), Natural Science Foundation of Heilongjiang Province of China (QC2015083), and Heilongjiang Postdoctoral Financial Assistance (LBH-Z14121).

Authors' contributions
The authors declare equal contribution. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]The Higher Educational Key Laboratory for Measuring and Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, 150080 Harbin, China. [2]School of Engineering, Harbin University, 150080 Harbin, China.

Received: 29 June 2016 Accepted: 14 February 2017
Published online: 23 February 2017

References

1. D Nister, H Stewenius, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Scalable recognition with a vocabulary tree, vol. 2 (IEEE, New York, 2006), pp. 2161–2168
2. Y Li, DJ Crandall, DP Huttenlocher, in *IEEE 12th International Conference on Computer Vision*. Landmark classification in large-scale image collections (IEEE, Kyoto, 2009), pp. 1957–1964
3. Y LeCun, L Bottou, Y Bengio, P Haffner, Gradient-based learning applied to document recognition. Proc. IEEE. **86**(11), 2278–2324 (1998)
4. S Haykin, *Neural networks and learning machines*, vol. 3. (Pearson, NJ, 2009)
5. A Krizhevsky, I Sutskever, GE Hinton, in *Advances in neural information processing systems*. Imagenet classification with deep convolutional neural networks (Neural Information Processing Systems (NIPS), Lake Tahoe, 2012), pp. 1097–1105

6. AV Singh, Content-based image retrieval using deep learning. Thesis, Rochester Institute of Technology. (2015)
7. V-A Nguyen, MN Do, in *IEEE International Conference on Multimedia and Expo (ICME)*. Deep learning based supervised hashing for efficient image retrieval (IEEE, Seattle, 2016), pp. 1–6
8. A Gordo, J Almazan, J Revaud, D Larlus, Deep image retrieval: learning global representations for image search (2016). arXiv preprint arXiv:1604.01325
9. B Girod, V Chandrasekhar, DM Chen, *et al*, Mobile visual search. IEEE Signal Proc. Mag. **28**(4), 61–76 (2011)
10. Z Li, K-H Yap, Content and context boosting for mobile landmark recognition. IEEE Signal Process. Lett. **19**(8), 459–462 (2012)
11. K-H Yap, Z Li, D-J Zhang, Z-K Ng, in *Proceedings of the 20th ACM International Conference on Multimedia*. Efficient mobile landmark recognition based on saliency-aware scalable vocabulary tree (ACM, Nara, 2012), pp. 1001–1004
12. Z Li, K-H Yap, An efficient approach for scene categorization based on discriminative codebook learning in bag-of-words framework. Image Vision Comput. **31**(10), 748–755 (2013)
13. Z Li, K-H Yap, Context-aware discriminative vocabulary tree learning for mobile landmark recognition. Dig. Signal Process. **24**, 124–134 (2014)
14. DP Huttenlocher, G Klanderman, WJ Rucklidge, *et al*, Comparing images using the Hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **15**(9), 850–863 (1993)
15. W Rucklidge, *Efficient visual recognition using the Hausdorff distance*, vol. 1173. (Springer-Verlag, Secaucus, 1996)
16. O Jesorsky, KJ Kirchberg, RW Frischholz, in *Audio-and video-based biometric person authentication*. Robust face detection using the Hausdorff distance (Springer, Berlin, 2001), pp. 90–95
17. DG Lowe, in *The Proceedings of the Seventh IEEE International Conference on Computer Vision*. Object recognition from local scale-invariant features, vol. 2 (IEEE, Kerkyra, 1999), pp. 1150–1157
18. DG Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
19. TJ Chin, H Goh, JH Lim, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Boosting descriptors condensed from video sequences for place recognition (IEEE, Anchorage, 2008), pp. 1–8
20. A Pronobis, B Caputo, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Confidence-based cue integration for visual place recognition (IEEE, San Diego, 2007), pp. 2394–2401
21. A Qamra, EY Chang, Scalable landmark recognition using extent. Multimedia Tools Appl. **38**(2), 187–208 (2008)
22. K Mikolajczyk, C Schmid, A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1615–1630 (2005)
23. L Van Gool, T Moons, D Ungureanu, Affine/photometric invariants for planar intensity patterns. Eur. Conf. Comput. Vis. **1064**, 642–651 (1996)
24. S Belongie, G Mori, J Malik, Matching with shape contexts. Stat. Anal. Shapes Part Ser. Model. Simul. Sci. Eng. Technol, 81–105 (2006)
25. L Bo, X Ren, D Fox, in *Advances in neural information processing systems*. Hierarchical matching pursuit for image classification: architecture and fast algorithms (Neural Information Processing Systems (NIPS), Granada, 2011), pp. 2115–2123
26. P Dreuw, P Steingrube, H Hanselmann, H Ney, G Aachen, in *British Machine Vision Conference*. Surf-face: face recognition under viewpoint consistency constraints (BMVA Press, London, 2009), pp. 1–11
27. S Lazebnik, C Schmid, J Ponce, in *IEEE Conference on Computer Vision and Pattern Recognition*. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, vol. 2 (IEEE, New York, 2006), pp. 2169–2178
28. DM Chen, G Baatz, K Koser, *et al*, in *IEEE Conference on Computer Vision and Pattern Recognition*. City-scale landmark identification on mobile devices (IEEE, Colorado, 2011), pp. 737–744
29. G Baatz, K Köser, D Chen, R Grzeszczuk, M Pollefeys, Leveraging 3d city models for rotation invariant place-of-interest recognition. Int. J. Comput. Vis. **96**(3), 315–334 (2012)
30. G Baatz, K Köser, D Chen, R Grzeszczuk, M Pollefeys, in *European Conference on Computer Vision*. Handling urban location recognition as a 2d homothetic problem (Springer, Crete, 2010), pp. 266–279
31. J Zhang, M Marszalek, S Lazebnik, C Schmid, in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*. Local features and kernels for classification of texture and object categories: a comprehensive study (IEEE, New York, 2006), pp. 13–13
32. E Tola, V Lepetit, P Fua, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. A fast local descriptor for dense matching (IEEE, Anchorage, 2008), pp. 1–8
33. E Tola, V Lepetit, P Fua, Daisy: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. **32**(5), 815–830 (2010)
34. N Khan, B McCane, S Mills, Better than sift? Mach. Vis. Appl. **26**(6), 819–836 (2015)
35. R Ji, LY Duan, J Chen, *et al*, in *18th IEEE International Conference on Image Processing*. Pkubench: a context rich mobile visual search benchmark (IEEE, Brussels, 2011), pp. 2545–2548
36. DW Scot, *Multivariate density estimation*. (Wiley & Sons, New York, 1992)
37. L Greengard, X Sun, A new version of the fast gauss transform. Doc. Math. **3**, 575–584 (1998)
38. K Sentz, S Ferson, *Combination of evidence in Dempster-Shafer theory*, vol. 4015. (Sandia National Laboratories, Albuquerque, 2002)
39. J Philbin, A Zisserman, The Oxford Buildings Dataset. (2007). http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/. Accessed 18 Feb 2017
40. MJ Swain, DH Ballard, Color indexing. Int. J. Comput. Vis. **7**(1), 11–32 (1991)
41. Y Liu, YF Zheng, in *IEEE International Joint Conference on Neural Networks*. One-against-all multi-class svm classification using reliability measures, vol. 2 (IEEE, Montreal, 2005), pp. 849–854
42. JC van Gemert, CJ Veenman, *et al*, Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1271–1283 (2010)
43. A Bosch, A Zisserman, X Muoz, Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal. Mach. Intell. **30**(4), 712–727 (2008)
44. Y Rubner, C Tomasi, LJ Guibas, in *Sixth International Conference on Computer Vision (ICCV)*. A metric for distributions with applications to image databases (IEEE, Bombay, 1998), pp. 59–66