Image Retrieval Using Random Forest-Based Semantic Similarity Measures and SURF-Based Visual Words

Anindita Mukherjee, Jaya Sil and Ananda S. Chowdhury

Abstract In this paper, we propose a novel image retrieval scheme using random forest-based semantic similarity measures and SURF-based bag of visual words. A patch-based representation for the images is carried out with SURF-based bag of visual words. A random forest, which is an ensemble of randomized decision trees, is applied next on a set of training images. The training images accumulate into different leaf nodes in each decision tree of the random forest as a result. During retrieval, a query image, represented using SURF-based bag of visual words, is passed through each decision tree. We define a query path and a semantic neighbor set for such query images in all the decision trees. Different measures of semantic image similarity are derived by exploring the characteristics of query paths and semantic neighbor sets. Experimental results on the publicly available COIL-100 image database clearly demonstrate the superior performance of the proposed content-based image retrieval (CBIR) method with these new measures over some of the similar existing approaches.

Keywords Semantic similarity measures · Random forest · Query path
SURF · Visual words

1 Introduction

Content-based image retrieval (CBIR) has emerged over the years as a popular area of interest for researchers in the computer vision and the multimedia communities. The principal aim of CBIR is to organize digital picture archives from a thorough

A. Mukherjee
Dream Institute of Technology, Kolkata, India

J. Sil
IIEST Shibpur, Howrah, India

A. S. Chowdhury (✉)
Jadavpur University, Kolkata, India
e-mail: aschowdhury@etce.jdvu.ac.in

© Springer Nature Singapore Pte Ltd. 2018

B. B. Chaudhuri et al. (eds.), *Proceedings of 2nd International Conference on Computer Vision & Image Processing*, Advances in Intelligent Systems and Computing 703, https://doi.org/10.1007/978-981-10-7895-8_7

analysis of their visual content [1]. Bag of visual words (BoVW) framework has become popular for modeling the image content [2]. In this model, an image is represented as a collection of elementary local features like SURF [3] or SIFT [4]. These local descriptors are then quantized by k-means algorithm to build a bag of visual words. In recent past, there have been efforts to improve the BoVW model. For example in [5], Bouachir et al. have used a fuzzy c-means-based approach to improve the retrieval performance. The authors in [6] have developed an affinity-based visual word assignment model. They have also proposed a new measure of dissimilarity using a penalty function. For probabilistic similarity measures in image retrieval, please see [7]. While these methods have shown promises, still there remains a wide scope to better the retrieval performance. One factor which has highly contributed to this scope is lack of proper semantic similarity measures. Notion of semantic similarity plays a pivotal role in the image content modeling and retrieval [8]. We have come across interesting works, where initial notions of semantic image similarity based on random forest are developed [9–11]. In this paper, we propose an image retrieval scheme using random forest-based semantic similarity measures and SURF-based BoVW model. The rationale behind using SURF features is its much faster execution time as compared to that of SIFT [3]. The main contribution of this paper is the design and detailed analysis of random forest-based new semantic similarity measures. Experimental comparisons on the publicly available COIL-100 [12] image database clearly show the merit of our approach.

2 Proposed Method

In this section, we describe in detail the proposed method. The section contains three parts. In the first part, we discuss the SURF-based BoVW model for image representation. We then describe how random forest is used for training. Finally, we derive novel semantic similarity measures based on random forest.

2.1 Image Representation Using SURF-Based BoVW

We first discuss basics of SURF features following [3]. We then mention how patch-based image representation is done using SURF-based BoVW. SURF uses Hessian Matrix to detect interest points. The Hessian Matrix $H(\mathbf{x}, \sigma)$ for any point $\mathbf{x} = (x, y)$ in an image I at a scale σ is mathematically expressed as:

$$H(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{yx}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix} \tag{1}$$

In Eq. (1), $L_{xx}(\mathbf{x}, \sigma)$ marks the convolution of the Gaussian second-order derivative $\frac{\delta^2}{\delta x^2} g(\sigma)$ with the image I at point $\mathbf{x}$ and so on. Integral images are used to efficiently obtain these computationally intensive convolutions. The above interest points are found across different scales (σ values). For the extraction of interest point descriptors (representation of neighborhood of any interest point, i.e., a patch), SURF uses sum of Haar wavelet responses. In the present problem, 64-dimensional SURF vectors are used to represent several patches in each training image. These local SURF descriptors need to be quantized to build the visual vocabulary. We apply k-means algorithm (with $k = 500$) to achieve this goal. Each cluster is treated as a unique visual word, and the collection of such visual words form the visual vocabulary [2]. Each image is then represented using a histogram of these visual words. Thus, at the end of this step, we have a 500-dimensional BoVW vector representing each training image.

2.2 Random Forest-Based Training

Here, we discuss how random forest can be used for training in this context of image retrieval. The rationale behind the choice of random forest is its very high accuracy and capability to handle large volume of data. Random forest is an ensemble classifier of decision trees with bagging (randomizing the training set) capability [13]. It votes for the most popular class among the individual trees. The information gain I for the jth node in a decision tree is given by:

$$I = H(S_j) - \sum_{i=L,R} \frac{|S_j^i|}{|S_j|} H(S_j^i) \tag{2}$$

In Eq. (2), $H(S)$ denotes the entropy of a node S, which for a discrete set of C labels is given by: $-\sum_{c \in C} p(c) log_2(p(c))$ and $|S_j|$ denotes the number of training images in the node S_j. So, $|S_j^L|$ and $|S_j^R|$, respectively, represent the number of training images in the left child and the right child of the node S_j. In this problem, we use the 500-dimensional BoVW vector and the class label for each training image as the two inputs to the random forest. At the end of this training phase, the training images are grouped into various leaf nodes in different decision trees.

2.3 Random Forest-Based Semantic Similarity Measures

Though random forest is mostly applied for classification, following [9], we have used it here to derive measures of semantic image similarity. In this section, we discuss three such measures. During the retrieval stage, a query image passes through each decision tree. Let m denote a training image, q denote a query image, and

t denote a decision tree in a random forest. Further, let M and T, respectively, denote the total number of training images and the total number of decision trees in the random forest. We now have the following definitions and expressions.

Definition 1 The **semantic neighbor set** $SNS(q,t)$ is defined following [9] as the set of training images present at the leaf node into which a query image q falls in a decision tree t.

Definition 2 The frequency-based similarity measure **sm1(m, q)** is defined as the number of trees ($t, t \in [1,T]$) in the random forest a training image m appears in $SNS(q,t)$. So, we mathematically express $sm1(m,q)$ as:

$$sm1(m,q) = \sum_{t=1}^{T} \phi_m(t) \tag{3}$$

Here, $\phi_m(t) = 1$ if $m \in SNS(q,t)$ and is 0 otherwise ($1 \leq m \leq M$).

Note that since $sm1$ is based on frequency, we do not normalize it.

Definition 3 A **query path** $p_k(q,t)$ of length k for a query image q in a tree $t(1 \leq t \leq T)$ is denoted by a sequence of nodes $n_0(t), n_1(t), \ldots, n_{(k-1)}(t)$, where $n_0(t)$ is the root, $n_i(t)$ is the ith intermediate node ($1 \leq i \leq (k-2)$), and $n_{(k-1)}(t)$ is a leaf node in the tree t. Here, q falls into $n_{(k-1)}(t)$ and the training images which are present in $n_{(k-1)}(t)$ form $SNS(q,t)$.

In Fig. 1, we show a typical query path in a decision tree using a sequence of red lines. The red oval marks the leaf node where the query image falls in this tree and xi in the same figure denotes the ith ($1 \leq i \leq 500$) element of the 500-dimensional BoVW vector. The label 5 in the leaf node indicates that the probability of class 5 is maximum at this node. However, we have determined the SNS, i.e., training images of different classes (and not just that of the highest class) which have accumulated in such nodes for the evaluation of our proposed semantic measures.

Definition 4 Let the set of $(k-1)$ features on the query path $p_k(q,t)$ be denoted by $f(t) = \{f_1(t), f_2(t), \ldots, f_{(k-1)}(t)\}$ and the set of weights (relative importance) of these k features be denoted by $\alpha(t) = \{\alpha_1(t), \alpha_2(t), \ldots, \alpha_{(k-1)}(t)\}$. Here, f_i connects $(n_i, n_{(i+1)}), (1 \leq i \leq (k-1))$, and so on. A query path-based similarity measure **sm2(m, q)** between a training image m and a query image q is defined as the summation over all trees the product of weights of all features appearing in a path in each tree. We mathematically express $sm2(m,q)$ as:

$$sm2(m,q) = \sum_{t=1}^{T} \prod_{i=1}^{(k-1)} \alpha_i(t) \tag{4}$$

We have actually used a normalized version of $sm2(m,q)$, defined as $sm2(m,q) = sm2(m,q)/max_{\forall m, 1 \leq m \leq M} sm2(m,q)$.

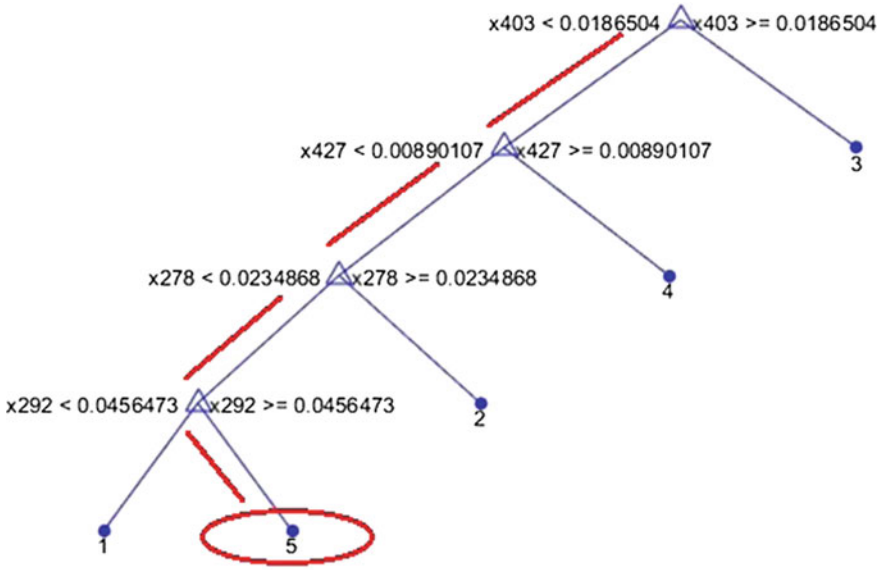

Fig. 1 A typical query path in a decision tree

Definition 5 Further, let the set of k features $f_i(t), (1 \leq i \leq (k-1))$ on the query path $p_k(q,t)$ be at respective levels $l_i(t), (1 \leq i \leq (k-1))$. Since each tree is essentially a binary tree, we define another query path-based similarity measure **sm3(m, q)** between a training image m and a query image q as the summation over all trees the product of level modulated weights of all features appearing in a path in each tree. So, $sm3(m,q)$ can be mathematically expressed as:

$$sm3(m,q) = \sum_{t=1}^{T} \prod_{i=1}^{(k-1)} \alpha_i(t) \times \frac{1}{2^{l_i(t)}} \tag{5}$$

We have actually used a normalized version of $sm3(m,q)$, defined as $sm3(m,q) = sm3(m,q)/max_{\forall m, 1 \leq m \leq M} sm3(m,q)$, where M denotes the total number of training images.

3 Complexity Analysis

In this section, we analyze the complexity of the construction of the random forest classifier and the computation of the three semantic measures. Let M, $|B|$, and $d(t)$ denote the number of training images, number of elements in the BoVW vector representing an image, and depth of a decision tree t. Note that in a random forest, as

a result of bagging, each decision tree is constructed with a randomly chosen subset of the number of elements in the BoVW vector ($f(t) = \{f_1(t), f_2(t), \ldots, f_{(k-1)}(t)\}$, $(k-1) \leq |B|$). At each node in a tree t, we have to compute the information gain for different $f_i(t)'s, (1 \leq i \leq (k-1))$. Then, the maximum cost of constructing a decision tree becomes $O(M|B|d(t))$. With a total of T such decision trees in the random forest and a maximum depth D, where $D = max_{\forall t, 1 \leq t \leq T}(d(t))$, the maximum cost of constructing the random forest is $O(M|B|DT)$. The maximum value of the length k of a query path $p_k(t)$ in a decision tree t is $d(t)$, which in the worst case can be D. So, the worst-case complexity for evaluating $sm1(m, q)$ (please see the definition in Eq. (3)) is $O(TD)$. The cost of using the weights $\alpha_i(t)$ for corresponding features $f_i(t), (1 \leq i \leq (k-1))$ in a decision tree t is $O(1)$. Note that these weights are already computed using Eq. (2) at the time of the construction of the decision trees. In the worst case, we need to use this $(D-1)$ times for all T trees. So, the worst-case complexity for evaluating $sm2(m, q)$ (please see the definition in Eq. (4)) is $O(T(D-1)) \approx O(TD)$. Similarly, the cost of evaluation of the levels $l_i(t)$ for any feature $f_i(t)$ in a decision tree t is also $O(1)$. In the worst case, we need to evaluate this $(D-1)$ times. So, the worst-case complexity for evaluating $sm3(m, q)$ (please see the definition in Eq. (5)) is $O(T(D-1)) + O(T(D-1)) = O(T(D-1)) \approx O(TD)$. The overall worst-case complexity of construction of the random forest and evaluation of any semantic measure is $O(M|B|DT) + O(TD) = O(M|B|DT)$.

4 Experimental Results

We use the publicly available COIL-100 image database [12] for experimentation. The database contains a total of 7200 images with 72 different images of 100 different objects having a viewpoint separation of 5°. We have used MATLAB as the computing platform. Precision and recall values are chosen as the measures of retrieval performance [1]. Precision indicates the percentage of retrieved images that are relevant to the query. In contrast, recall measures the percentage of all the relevant images in the database which are retrieved. Precision versus recall curves are obtained by changing the thresholds θ_1, θ_2, and θ_3 in connection with the three semantic similarity measures $sm1, sm2$, and $sm3$, respectively. So, for obtaining the curve using measure $sm1$, we vary θ_1 from 5 to the total number of decision trees in the random forest and retrieve only those training image(s) m for which $sm1(m, q) > \theta_1$. Likewise, we obtain the curves using $sm2$ and $sm3$ by varying θ_2 and θ_3 from 0.05 to 1.0.

We experimentally determine the optimal number of trees in the random forest to be 100. Please see Fig. 2 where the best retrieval performance for Coil 1 is achieved with $T = 100$. We now compare our performance with a *fuzzy weighting scheme* [5], a *term frequency–inverse document frequency (tfx)*-based approach [14], a method which only uses *term frequency (txx)* [15], and a visual word assignment model (vwa) [6].

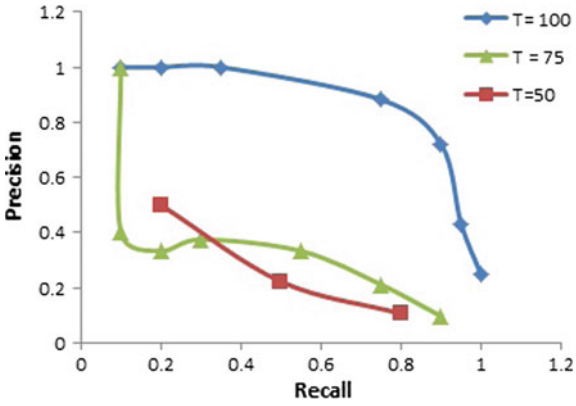

Fig. 2 Precision versus recall curves for Coil 1 using $sm1$ and three different values of number of decision trees ($T = 100, T = 75, T = 50$)

The precision versus recall curves are shown for two different query objects, namely Coil 3 and Coil 10 for each of the three semantic measures $sm1$, $sm2$, and $sm3$ are shown in Fig. 3 and Fig. 4, respectively. The curves clearly indicate that the retrieval performance using all three proposed semantic similarity measures yields superior results compared to the four competing methods. We now show the retrieved

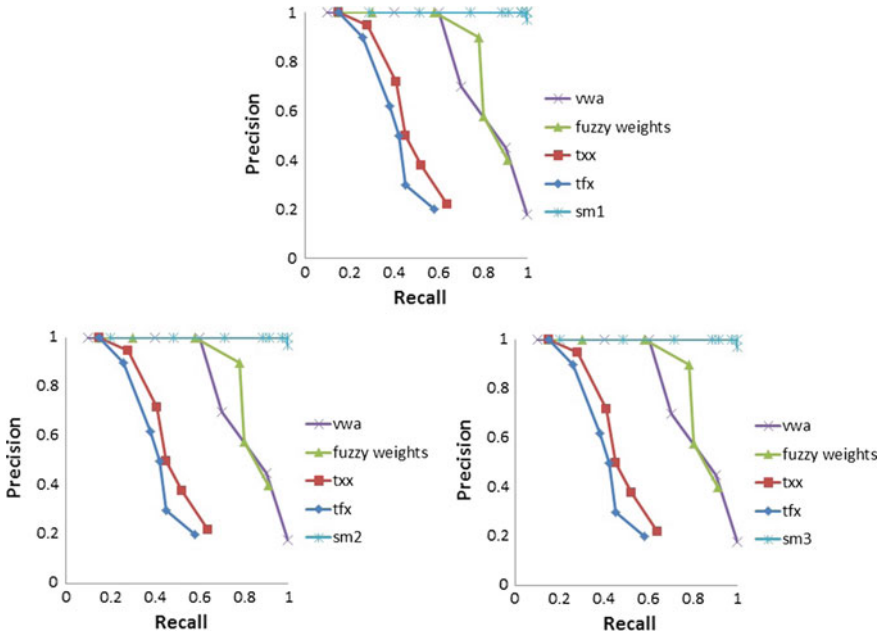

Fig. 3 Precision versus recall curves of five different methods, namely txx [15], tfx [14], fuzzy weighting [5], vwa [6], and current approach for Coil 3 with three different semantic similarity measures: $sm1$ (top), $sm2$ (bottom left), $sm3$ (bottom right)

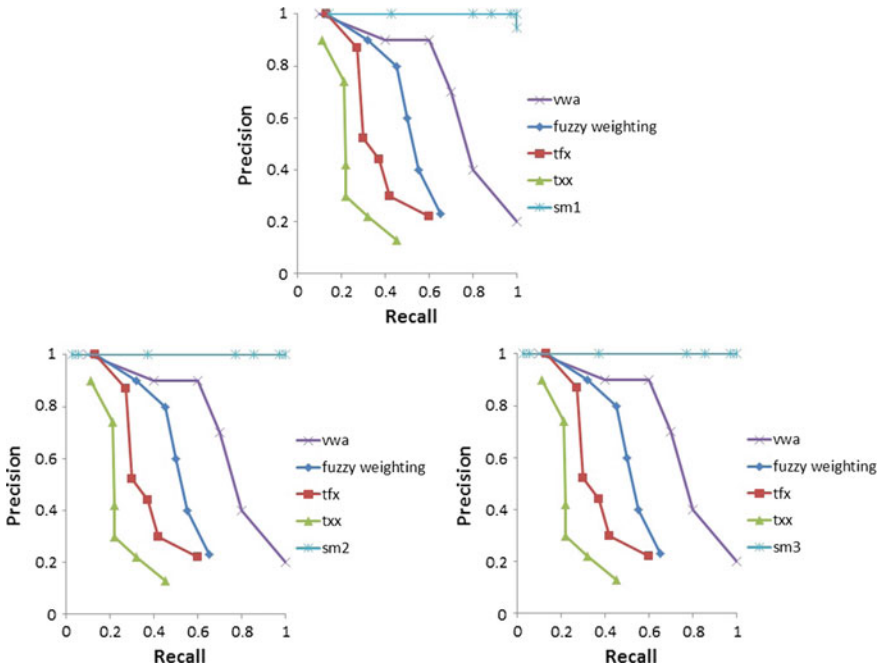

Fig. 4 Precision versus recall curves of five different methods, namely txx [15], tfx [14], fuzzy weighting [5], vwa [6], and current approach for Coil 10 with three different semantic similarity measures: *sm*1 (top), *sm*2 (bottom left), *sm*3 (bottom right)

images for Coil 3 and Coil 10 in Figs. 5 and 6. The retrieved results illustrate that the top five retrieved images for all three semantic measures are relevant. The rank and set of the relevant images are, however, different for different measures. Now, we include a failed case for the object Coil 9 in Fig. 7. This figure indicates that all three measures fail to retrieve only relevant images (images belong to the same class as that of the query image). The reason for failure is that there are quite a few extremely similar objects like Coil 9 in the database. Still, the measures *sm*2 and *sm*3 yield better results than *sm*1. This is because *sm*1 is only based on frequency of appearance of a training image in the SNS of a query image. In contrast, both *sm*2 and *sm*3 are derived from the characteristics (weights and levels of BoVW elements) of a query path.

We also present the recognition rate as an average precision for ten different objects, namely Coil 1 to Coil 10 of the COIL-100 database [5]. Please note that the recognition rate does not take into account any recall. In Table 1, we compare the recognition rates for the above objects of the proposed method with three different semantic similarity measures against four competing methods. The results clearly demonstrate that all three measures in our method have better performances than the competing methods. In nine out of ten cases, it turns out that the three measures become (single or joint) winners having achieved the highest recognition rate.

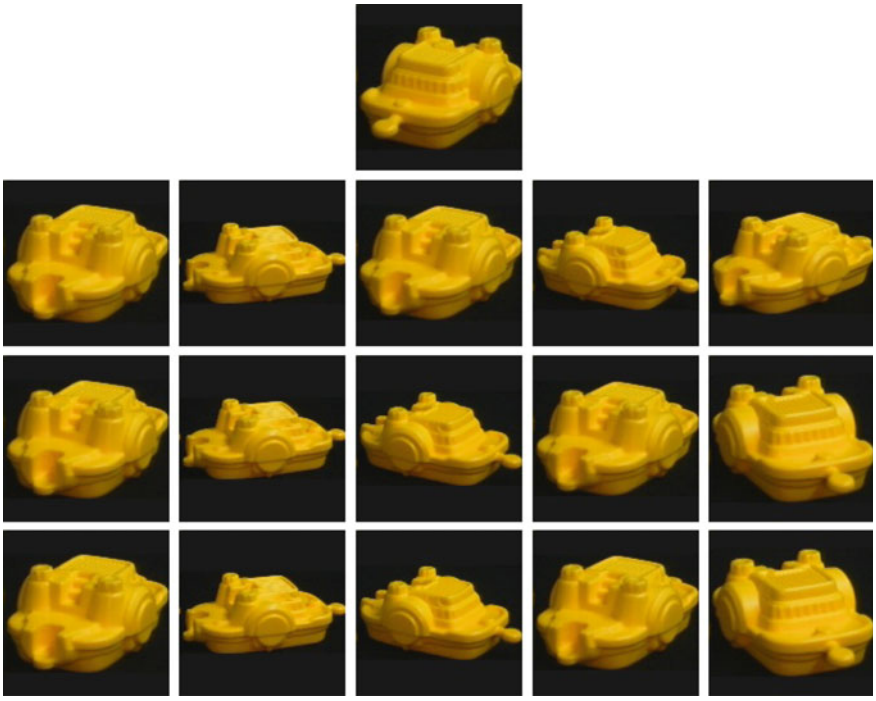

Fig. 5 Retrieval results for COIL-3: query image (first row), top five retrieved images based on *sm*1 (second row), *sm*2 (third row), and *sm*3 (fourth row). All five retrieved images for all three measures are relevant (belong to the same class as that of the query image)

Furthermore, all three average recognition rates, namely 89.7% from *sm*1, 93.2% from *sm*2, and 92.1% from *sm*3, clearly surpass the previously reported recognition rates. Once again, (and in fact, generally speaking), among the proposed three measures, *sm*2 and *sm*3, which carry more information, yield better results than *sm*1. Fourth best is [6] with an average recognition rate of 86%, followed by [5] with the reported average recognition rate of 80%. The other two methods [14, 15] are clearly quite behind with average recognition rates of 71.5% and 61.5%, respectively.

5 Conclusion

In this paper, we proposed a method of image retrieval using random forest-based new semantic similarity measures and SURF-based bag of visual words. The semantic similarity measures are derived from characterization of query paths and semantic neighbor sets in each decision tree of the random forest. Comparisons with some of the existing approaches on the COIL-100 database clearly show the merits of the proposed formulation. In future, we plan to perform more experiments with other similar

Fig. 6 Retrieval results for COIL-10: query image (first row), top five retrieved images based on *sm*1 (second row), *sm*2 (third row), and *sm*3 (fourth row). All five retrieved images for all three measures are relevant (belong to the same class as that of the query image)

Table 1 Recognition rate comparison among different competing methods: txx [15], tfx [14], fuzzy weighting [5], vwa [6], and current method with semantic similarity measures *sm*1, *sm*2, and *sm*3

Image	txx	tfx	Fuzzy weighting	vwa	sm1	sm2	sm3
Coil 1	0.5	0.4	0.65	0.8	0.98	0.98	0.98
Coil 2	0.4	0.1	0.45	0.6	0.88	0.89	0.89
Coil 3	0.9	0.95	1.0	1.0	1.0	1.0	1.0
Coil 4	1.0	0.9	1.0	1.0	1.0	1.0	1.0
Coil 5	0.25	0.1	0.75	0.75	0.97	0.98	0.98
Coil 6	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Coil 7	1.0	0.85	0.95	1.0	0.94	0.93	0.91
Coil 8	0.55	0.5	0.7	0.75	0.94	0.94	0.94
Coil 9	0.7	0.6	0.6	0.7	0.26	0.6	0.51
Coil 10	0.85	0.75	0.9	1.0	1.0	1.0	1.0
Average	0.715	0.615	0.8	0.86	0.897	0.932	0.921

Fig. 7 Retrieval results for COIL-9 showing some failed cases: query image (first row), top five retrieved images based on sm1; only fifth image is from the relevant class (second row), top five retrieved images based on sm2; first, fourth, and fifth from relevant classes (third row), top five retrieved images based on sm3; first, fourth, and fifth from relevant classes (fourth row)

approaches on additional databases like Oxford buildings [11]. We will also exploit contextual and structural information in random forests [16] as well as explore deep learning-based approaches [17] to further improve the retrieval performance.

References

1. Datta, R., Joshi, D., Li, J., Wang, James Z., Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys, 40(2), 1–60, (2008).
2. Sivic, J., Zisserman, A.: Video Google: Efficient Visual Search of Videos, In Toward Category-Level Object Recognition, 127–144, (2006).
3. Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L.: Speeded-up robust features (SURF), Computer Vision and Image Understanding, 110(3), 346–359, (2008).
4. Lowe D. G.: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60(2), 91–110, (2004).
5. Bouachir, W., Kardouchi, M., Belacel, N.: Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting Scheme for Efficient Indexation, Proc. SITIS, 215–220, (2009).

6. Mukherjee, A., Chakraborty, S., Sil, J., Chowdhury, A.S.: A Novel Visual Word Assignment Model for Content Based Image Retrieval, Proc. CVIP, Balasubramanian Raman et al. (eds.), Springer AISC, Vol. 459, 79–87, (2016).
7. Rahman, M.M., Bhattacharya, P., Kamel, M., Campilho A.: Probabilistic Similarity Measures in Image Databases with SVM Based Categorization and Relevance Feedback, Proc. ICIAR, Springer LNCS, Vol. 3656, 601–608, (2005).
8. Liu Y., Zhang D., Lu G., Ma W-Y.: A survey of content-based image retrieval with high-level semantics, Pattern Recognition, 40(1), 262–282, (2007).
9. Fu, H., Qiu G.: Fast Semantic Image Retrieval Based on Random Forest, Proc. ACM MM, 909–912, (2012).
10. Moosman, F., Triggs, B. and Jurie, F.: Fast Discriminative Visual Codebooks using Randomized Clustering Forests, Proc. NIPS, 985–992, (2006).
11. Dimitrovski, I., Kocev, D., Loskovska, S., Dzeroski, S.: Improving bag-of-visual-words image retrieval with predictive clustering trees, Information Science, 329(2), 851–865, (2016).
12. Nene, S. A., Nayar, S. K., Murase, H.: Columbia Object Image Library (COIL-100), Tech. Report, Department of Computer Science, Columbia University CUCS-006–96, (1996).
13. Breiman, L.: Random Forests, Machine Learning, 45, 5–32, (2001).
14. Sivic, J., Zisserman A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, Proc. ICCV, 470–1477, (2003).
15. Newsam, S., Yang Y.: Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery, Proc. ACM GIS, Article No. 9, (2007).
16. Kontschieder P., Rota Bulo S., Pelillo M.: Semantic Labeling and Object Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(10), 2104–2116, (2014).
17. Wan J. et al.: Deep Learning for Content-Based Image Retrieval: A Comprehensive Study, Proc. ACM MM, 157–166, (2014).