Feature Weighting in Visual Product Recognition

Wen Zhang, Kim-Hui Yap, Da-Jiang Zhang and Zhen-Wei Miao
School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 639798
Email: {WZHANG017, EKHYAP, DZHANG3, ZWMIAO}@ntu.edu.sg

***Abstract*—Significant progress towards visual search has been made in the past two decades through the development of local invariant features. Among existing local feature detectors, the Scale Invariant Feature Transform (SIFT) is widely used since it is designed to be invariant to minimal illumination changes and certain geometric transformations. However, in practice, the recognition performance is still subject to actual condition. Some keypoints are more stable while others are less stable and can not be repeatedly detected. Besides, in visual object recognition where the foreground object is to be recognized while the background suppressed, the current scalable vocabulary tree (SVT) framework treats each descriptor as equally important, hence restricting its performance. This paper aims to study the effect of SIFT respect to illumination and geometric changes and develop a feature weighting algorithm to incorporate the stability of SIFT and saliency information into weighted scalable vocabulary tree (WSVT) based recognition. Experimental results on a commercial product database show the proposed feature weighting algorithm outperforms the baseline SVT recognition by 5%.**

I. Introduction

In recent years, local feature detection has drawn great attentions and made a tremendous impact on visual search. Local features have many advantages over global features, such as sparse representation of the image, robustness to occlusion and noisy background. Many works have been done to comprehensively evaluate local feature detectors using various criteria, including visual inspection, repeatability, consistency and matching score [1], [2]. Among various local feature detectors, scale-invariant feature transform (SIFT) detector, proposed by Lowe [3] is one of the state-of-art algorithms. It extracts local features by employing the difference of Gaussian to approximate the normalized Laplacian of Gaussian of the image. Compared with other interest point detectors, this method speeds up the process and improve the computational efficiency.

The local features extracted by the SIFT algorithms provide an efficient way for image recognition and retrieval. However, directly computing the distance between descriptors is very time consuming. Bag-of-Words (BoW) model [4] is proposed to address the issue and improve the image recognition efficiency. After constructing a codebook by applying K-means clustering, each image is summarized as a Bag-of-Words (BoW) histogram according to its distribution of word occurrence. In order to enable the use of a larger vocabulary, an approach called scalable vocabulary tree (SVT) [5] combined with inverted index was proposed. The SVT for a particular image database is built by recursively applying hierarchical k-means clustering on all the feature descriptors representing the training database. When performing image matching, the feature descriptors of the query image are passed through the vocabulary tree and represented as a histogram by applying a hierarchical scoring strategy. The database images are then quickly scored accordingly.

In spite of the notable strides achieved within the last decades, matching and learning visual objects is still challenging on a number of aspects. First, images from the same object category can produce very different images in diverse lighting environment, perspective changes, partial occlusions, and cluttered backgrounds. SIFT is designed to handle a certain amount of illumination and geometric transformation. However, in practice, the recognition performance is still subject to actual condition. There still exist some cases that SIFT can not handle, such as non-linear illumination changes [6] and certain affine transformation [7]. Some keypoints are more stable, while others are less stable and can not be repeatedly detected in these circumstances. Second, in the standard SVT-based image recognition, the local descriptors in the whole image are treated with equal importance. This restricts its performance as in most scenarios the local descriptors on the foreground object should be given more weights than the descriptors in the background.

In view of this, we propose a new feature weighting algorithm: (1) By studying the effect of SIFT with respect to illumination and geometric variations, different scores are assigned to interest points according to their stability in the augmented samples of various transformations. Hence the stable interest points can be obtained. (2) the saliency information is incorporated into feature weighting, where more salient descriptors on the foreground are strengthened with larger weights.

II. PROPOSED METHOD

In order to address the issue that SIFT is unstable under certain photometric and geometric transformations, feature weighting methods based on photometric and geometric transformations are introduced individually. Then the two feature weighting methods are combined together with saliency information to further improve the performance. The proposed image recognition framework is illustrated in Fig. 1.

A. Photometric transformation feature weighting

Although SIFT is designed to handle limited changes in illumination, it is observed that some SIFT descriptors are not stable in different lighting conditions. In view of this, in

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by a grant from the Singapore National Research Foundation and administered by the Interactive & Digital Media Programme Office at the Media Development Authority.

978-1-4799-8391-9/15/$31.00 ©2015 IEEE

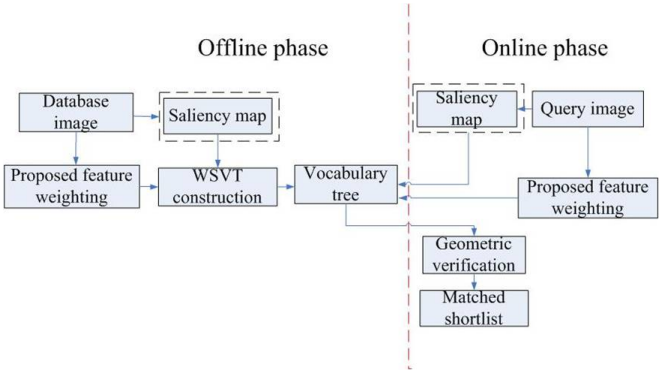

Fig. 1. Proposed image recognition framework

the photometric feature weighting method, the descriptors that are more stable under different photometric transformations are given more importance. There exists three main ways to adjust brightness and contrast of a digital image [8]. (1) Brightness adjustment. (2) Contrast of the middle-tone pixel value adjustment. (3) Gamma correction. Gamma correction is a nonlinear operation commonly used in photography to adjust the image contrast, let g be the normalized pixel value of an image, $0 \leq g \leq 1$:

$$G_2(g, \gamma) = g^{\frac{1}{\gamma}} \tag{1}$$

where γ is the gamma parameter. The process of utilizing $\gamma < 1$ is called gamma compression, which tends to make the image darker. On the contrary, the image is brightened when $\gamma > 1$ and the application of the expansive power-law nonlinearity is called gamma expansion. In our works we adopt gamma correction to artificially generate images to simulate different lighting environments.

The photometric transformation feature weighting includes:

1) Image generation and feature extraction. For each image in the database, we artificially generate p photometric-transformed images whose gamma values form a geometric sequence. Suitable gamma value settings are selected by testing on different ratio values of the geometric sequence and the value of p. Then we extract SIFT keypoints for all the $p + 1$ reference images.
2) Feature selection and weight assignment. We compare the similarity between each keypoint in the image and the keypoints in the p generated images. The keypoint in the image is considered as repeatedly detected and assigned weight w according to the number of repeatibility, $w \in \{1, 2, \cdots, p + 1\}$ if the similarity meets the following two requirements.

We define the similarity between the keypoint in the image and the keypoint in generated images according to two original requirements: (1) Neighborhood Checking. The spatial distance between two points is smaller or equal to 2. (2) Descriptor Similarity. $sim(A, B) = \mathbf{A} * \mathbf{B}/(|A| * |B|) \geq 0.94$, where A and B represent two descriptors. In the end, all the weights are normalized to be between 0-1.

B. Geometric transformation feature weighting

Although SIFT is designed to be invariant to certain amount of geometric transformation, the recognition performance is subject to actual conditions. Some SIFT keypoints are found to be stable in certain geometric transformations. In view of this, we developed a geometric transformation feature weighting strategy. By assigning higher weights to more stable keypoints, the recognition performance can be improved.

In order to check the stability of the keypoints under different geometric transformations, we first need to artificially generate some new images after various geometric transformations. Suppose the coordinate of a pixel in the original image is $[x, y]$, we can use a 2 by 2 matrix to perform the transformation:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{2}$$

where $[\hat{x}, \hat{y}]$ is the coordinate of the generated new pixel location. The four parameters, a_1, a_2, a_3 and a_4 control the affine transformation of the image. In our experiments, we use five types of affine transformations, including vertical & horizontal stretching, scaling, shearing and rotation. The geometric transformation feature weighting is given as follows:

1) Image generation. For each image, we artificially generate four different geometric-transformed images based on each transformation type. SIFT keypoints are extracted for the original image and the four generated images for each transformation type.
2) Keypoint back-projection. The coordinate of each SIFT keypoint is back-projected to the original image.
3) Feature selection and weight assignment. Compare the similarity between each keypoint in the image and the keypoints in four generated images after back-projection. If a keypoint is repeatedly detected in t generated images by checking the similarity criteria described in the last section, assigning weight w equal to $t + 1$ for all the $t + 1$ keypoints. All the weights $w \in \{1, 2, \cdots, 5\}$ are normalized to be between 0-1.
4) Transformations are then combined to further improve the recognition performance.

C. Transform and saliency feature weighting

In the previous two sections we introduce two transform feature weighting algorithms based on photometric and geometric transformation, respectively. In fact, the two algorithms can be combined to further improve the recognition rate. For geometric feature weighting, suppose finally g additional reference images are generated for each image after combination. The same feature weighting method is performed to the p photometric-transformed reference images and g geometric-transformed reference images. Suppose M descriptors are kept for all the reference images after feature selection and weighting. All the weights $w \in \{1, 2, \cdots, p + g + 1\}$ are normalized to be between 0-1.

1) Saliency map generation: Generally, the descriptors on the foreground objects should also be given higher importance than the background descriptors. Hence a saliency map is

produced for each image according to Graph-based Visual Saliency (GBVS) [9]. This approach includes two steps to generate a saliency map. Firstly, an activation map A is produced with the given the feature map F. In GBVS, an importance factor is assigned to the directed edge from location (i, j) to location (p, q):

$$e((i,j),(p,q)) = \left|\log \frac{F(i,j)}{F(p,q)}\right| * \exp\left(-\frac{(i-p)^2+(j-q)^2}{2\alpha^2}\right) \tag{3}$$

where α is a free parameter. Then we can get the activation map A according to eigenvector computation based on graph theory. Secondly, normalization followed by map combination are performed to obtain the final saliency map S of the given image. After saliency map generation, each local descriptor has its own saliency value in the saliency map, $S = \{s_i\}, i = 1, 2, \cdots, M$, normalized to be between 0-1.

2) WSVT construction: The traditional SVT treats each sample as equally important in K-means clustering. However, the algorithm can be enhanced to consider samples with different weights. Hence in place of traditional SVT, the weighted SVT (WSVT) is developed, where the obtained transform weights and saliency values are incorporated in cluster center calculation. Saliency values $S = \{s_1, s_2, \cdots, s_M\}$ and the transform weights $W = \{w_1, w_2, \cdots, w_M\}$ are multiplied to obtain the final weights. All of the obtained descriptors, $D = \{d_i\}, i = 1, 2, \cdots, M$, and their corresponding weights $W_f = \{s_i w_i\}, i = 1, 2, \cdots, M$ are used in WSVT construction. The WSVT algorithm are summarized in Algorithm 1.

Algorithm 1 WSVT construction

Input: SIFT descriptors $\mathbf{D} = \{d_1, d_2, \cdots, d_M\}$ and their corresponding weights $\mathbf{W_f} = \{s_1 w_1, s_2 w_2, \cdots, s_M w_M\}$;
Output: A vocabulary tree of L levels with branch factor K consisting of K^L leaf nodes, $n^{l,h} \in T$, where $l \in \{0, 1, \cdots, L\}$ indicates the level, $h \in \{1, 2, \cdots, K^{L-l}\}$ indicates the node's index at l^{th} level. Set the branch factor K = 10, the depth L = 5;
1: Start with $l = L$, applying K-means clustering to partition all samples into K^{L-l} clusters. Each cluster $n^{l,h}$ contains samples $D^{l,h}$, $h \in \{1, 2, \cdots, K^{L-l}\}$;
2: Update $n^{l,h}$, $h \in \{1, 2, \cdots, K^{L-l}\}$ using the weight information. $n^{l,h} \leftarrow \sum_{d_h \in D^{l,h}} s_i w_i d_i \;/\; \sum_{d_i \in D^{l,h}} s_i w_i$;
3: For each cluster $n^{l,h}$, generate K finer clusters for $(l-1)^{th}$ level: $\mathbf{N} = \{n^{l-1,(h-1)K+a}\}$, $a \in \{1, 2, \cdots, K\}$;
4: Repeat the pattern assignment and cluster center update until there's no change of the cluster centers for level l;
5: $l \leftarrow l - 1$, repeat the weighted K-means clustering until $l = 0$;

3) Image representation: For an image consisting of m descriptors $\mathbf{D} = \{d_i\}, i = 1, 2, \cdots, m$, each descriptor d_i is passed through the weighted vocabulary tree at each level to choose its closest clusters. As such d_i is quantized to $T(d_i) = \{n^{l,h_i}\}_{l=1}^{L}, h_i \in 1, 2, \cdots, K^{L-l}$. Then the entire image is summarized as a BoW histogram $\mathbf{H}$ based on the descriptor distribution on T. Each descriptor is assumed to have equal contribution in the traditional BoW histogram representation. However, in our proposed method, in order to enhance the descriptors that have high stability and saliency value, the corresponding weights $W_f = \{s_i w_i\}, i = 1, 2, \cdots, m$ of descriptors are incorporated in BoW histogram construction as follows:

$$\mathrm{H(c)} = \frac{m}{\sum_{i=1}^{m} s_i w_i} \sum_{i=1}^{m} s_i w_i \cdot \mathrm{I}\big(\mathrm{c} = \arg\min(||\mathrm{n} - \mathrm{d_i}||_2)\big)_{n \in T} \tag{4}$$

where I is the indicator function.

During online phase, the same feature weighting method is applied to each query image. The weighted BoW histogram for a query image is built by traversing through the WSVT. The similarity between a database image $\mathbf{H}_d$ and a query image $\mathbf{H}_q$ can be computed as eq. (5), where f is the inverted-index vector in [5]. The shortlisted database images of highest similarity based on histogram comparison will be selected.

$$sim(\mathbf{H}_q, \mathbf{H}_d) = \left|\left| \frac{\mathbf{H}_q \cdot \mathbf{f}}{||\mathbf{H}_q \cdot \mathbf{f}||} - \frac{\mathbf{H}_d \cdot \mathbf{f}}{||\mathbf{H}_d \cdot \mathbf{f}||} \right|\right| \tag{5}$$

Finally, Geometric verification (GV) [10] is applied to ensure the consistency of feature matches between the query image and database images. Based on the geometric transformation model estimated by random sample consensus (RANSAC), the plausibility of the matches is improved by rejecting inaccurate and inconsistent matches.

III. EXPERIMENTAL RESULTS

We conducted our proposed feature weighting image recognition strategy on a commercial product database. The database is consisted of 3882 reference images and 333 query images for 41 commercial product categories. Sample images are shown in Fig. 2. The commercial products cover a wide range of common commodities, including snacks and other daily supplies. The reference images are high quality images taken under good condition. On the contrary, the test images are manually captured using mobile phone in different conditions, including illumination variation, cluttered backgrounds, partial occlusion, viewpoints and scale.

(a)

(b)

Fig. 2. Sample images from the commercial product database: (a) Reference images; (b) Test images.

First, experiments on photometric and geometric feature weighting are performed individually. All of the results are

obtained after geometric verification. For photometric feature weighting, first we conducted some experiments to select appropriate parameters for the p artificially generated images whose gamma values form a geometric sequence. By changing the ratio value r of the geometric sequence and the value of p, different gamma value settings are tested. Experimental results show a high recognition rate of 0.892 can be achieved when $r = 2.1$ and $p = 5$, while the recognition rate for SVT method in [5] without feature weighing is 0.850. Therefore, it's evident that the proposed photometric feature weighting outperforms the baseline SVT by 4.2%.

For geometric feature weighting, first five types of geometric transformation are tested individually. The experiment results are shown in Table. I. The table shows that using scaling alone can offer the most improvement, followed by rotation, shearing, horizontal stretching, vertical stretching. Then we combine these five types gradually from the one with best improvement to the one with the lowest. The experimental results show that the combination of scaling and rotation can give the best improvement to achieve a recognition rate of 0.887. The performance does not show clear improvement as more transformation types are included. Compared with the baseline SVT, the proposed geometric feature weighting significantly improves the recognition rate by 3.7%.

TABLE I. EXPERIMENTAL RESULTS FOR INDIVIDUAL GEOMETRIC TRANSFORMATION

SVT baseline	0.850
Scaling	0.883
Rotation	0.881
Shearing	0.869
Horizontal stretching	0.859
Vertical stretching	0.846

Table. II shows the experimental results with respect to different feature weighting methods. (a) baseline SVT; (b) Photometric feature weighting; (c) geometric feature weighting; (d) combined photometric & geometric feature weighting; (e) transform and saliency feature weighting. From the table we can see (1) Both photometric and geometric feature weighting outperforms baseline SVT; (2) Combined photometric and geometric feature weighting improves the recognition rate to 0.897, better than performing the two feature weighting strategies individually; (2) By incorporating saliency information, the performance of transform feature weighting is further improved by 0.9%. A typical recognition result comparison of the SVT method [5] and the proposed method is shown in Fig. 3. Using the proposed feature weighting, the descriptors of high stability and saliency value are assigned higher importance, resulting in correct match. While in the other method, the best matched image is incorrect.

IV. CONCLUSION

A new transform and saliency feature weighting algorithm is proposed in this paper. By giving different weights to SIFT descriptors according to their stability in various transformation, we can address the problem that some SIFT descriptors are unstable due to illumination changes and certain geometric transformation. Besides, we also incorporate saliency information with the transform feature weighting algorithm to further improve the recognition rate. Experimental results on a commercial product database demonstrated that the recognition rate is significantly improved when compared with the baseline SVT recognition.

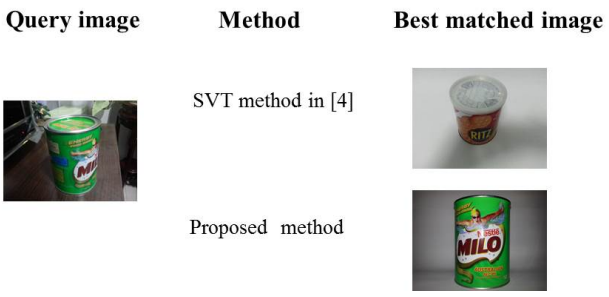

Fig. 3. Comparison of the best matched image of the SVT method and proposed feature weighting method

TABLE II. EXPERIMENTAL RESULTS FOR DIFFERENT FEATURE WEIGHTING METHODS

Photometric	Geometric	Saliency	Recognition rate
×	×	×	0.850 (baseline)
√	×	×	0.892
×	√	×	0.887
√	√	×	0.897
×	×	√	0.881
√	√	√	**0.906** (proposed)

REFERENCES

[1] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[2] Z. W. Miao and X. D. Jiang, "Interest point detection using rank order LoG filter," *Pattern Recognition*, vol. 46, no. 11, pp. 2890–2901, November 2013.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision,*, vol. 60, no. 2, pp. 91–110, 2004.

[4] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[5] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2161–2168.

[6] R. J. Alitappeh and F. Mahmoudi, "Mgs-sift: A new illumination invariant feature based on sift descriptor," *International Journal of Computer Theory & Engineering*, vol. 5, no. 1, 2013.

[7] J. M. Morel and G. S. Yu, "Asift: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438–469, 2009.

[8] J. Y. Fan, H. Cao, and A. C. Kot, "Estimating exif parameters based on noise features for image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 4, pp. 608–618, 2013.

[9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[10] B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, "Mobile visual search," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, 2011.