International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:6, 2017

Visual Thing Recognition with Binary Scale-Invariant Feature Transform and Support Vector Machine Classifiers Using Color Information

Wei-Jong Yang, Wei-Hau Du, Pau-Choo Chang, Jar-Ferr Yang, Pi-Hsia Hung

***Abstract*—**The demands of smart visual thing recognition in various devices have been increased rapidly for daily smart production, living and learning systems in recent years. This paper proposed a visual thing recognition system, which combines binary scale-invariant feature transform (SIFT), bag of words model (BoW), and support vector machine (SVM) by using color information. Since the traditional SIFT features and SVM classifiers only use the gray information, color information is still an important feature for visual thing recognition. With color-based SIFT features and SVM, we can discard unreliable matching pairs and increase the robustness of matching tasks. The experimental results show that the proposed object recognition system with color-assistant SIFT SVM classifier achieves higher recognition rate than that with the traditional gray SIFT and SVM classification in various situations.

***Keywords*—**Color moments, visual thing recognition system, SIFT, color SIFT.

I. Introduction

FOR smart production systems, it is desirable to build a general visual thing recognition system [1], [2] to assist the users to recognize the visual thing and retrieve its embedded information such as nutrition, ingredients, manufacture date, prices, etc. of the products at once. In literatures, there are many object recognition systems [3], [4] which have been proposed. However, deformations of visual thing for soft objects, arbitrary orientations, and various capturing distance degrade the performance of recognition systems dramatically in real applications. To cover all possible deformations, the scale SIFT features suggested by Lowe [5], [6] become a good selection to achieve robust visual thing recognition systems, where the SIFT features use Gaussian pyramids to compute multiscale differences to represent the local features. The speeded-up robust feature (SURF) [7] adopts a fast Hessian detector, integral image, and gradient-based descriptors to generate local features. Comparing to the SIFT, the SURF has lower time consumption by integral images. The other region-based features, such as the maximally stable extremal regions (MSER) [8] and histogram of oriented gradient (HOG) [9] are also good features. The other popular operators to detect the interest points and features could be Harris corner detector [10], Hough transform [11], Shi and Tomasi features [12]. It has been verified [13] that the performance of SIFT is better than that of SURF. Thus, in this paper, the SIFT is chosen as the kernel feature in our research because the SIFT feature presents its stability in most situations. Although the SIFT is slower in computation, the advances of CPU and GPU as well as the cloud computing actually can help to alleviate its computation problem. Although the SIFT is a robust feature descriptor, it still cannot separate the feature points, which are with different colors. To further include color information, the color-SIFT features are also proposed in [14].

To simplify cloud server systems, the compact descriptor for visual search (CDVS) [15], which is a binary SIFT description, has been proposed in MPEG standardization for visual search applications. The binary SIFT features will help to achieve lower data transmission bandwidth for client server cooperative visual search systems. In order to achieve a real-time visual thing recognition system, in this paper, a color-binary-SIFT descriptor is proposed to replace the original binary SIFT for visual thing recognition. The color information is considered because it provides valuable information in object description and matching tasks. Figs. 1 and 2 show examples for color influence among the product and flower objects. In the gray scale domain, it becomes difficult to differentiate the products, where different colors might represent different integrations. Without color information, it becomes hard to identify the correct one. It will cause confusion between two similar products. By properly including the color information, we could easily identify the differences for the detection. So, a new descriptor, called color-SIFT, is proposed in this paper to achieve more robust performance than the traditional SIFT.

W.-J. Yang, P.-C. Chang and J.-F. Yang are with the Institute of Computer and Communication Engineering, the Department of Electrical Engineering, the National Cheng Kung University, Tainan 701, Taiwan (phone: +886-916727713, e-mail: weijong@hotmail.com, pcchung@ee.ncku.edu.tw, jefyang@mail.ncku.edu.tw).

W.-H. Du was with the Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan (e-mail: vcxz80688@hotmail.com).

P.-H, Hung is with Department of Education, National University of Tainan, Tainan 700, Taiwan (e-mail: hungps@mail.nutn.edu.tw).

Fig. 1 Neglecting the color information may lose product distinction

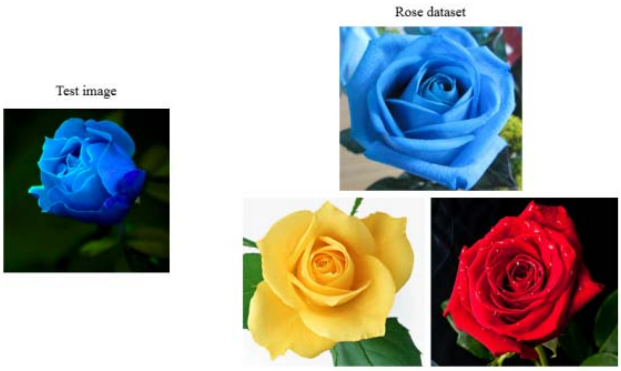

Fig. 2 Neglecting the color information may lose flower distinction

II. Traditional SIFT Features

The SIFT features [5] which were proposed by Lowe in 1999 describe the distinct local features in the image. The SIFT features can robustly identify the object even with clutter and under occlusion, because they are invariant to scaling, orientation, illumination changes, and partially distortion. In order to perform the scale invariant, the SIFT uses Gaussian filters with different blur parameters σ to generate many blurred images by

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (1)$$

where

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (2)$$

The difference of successive Gaussian-blurred images, called difference of Gaussian (DoG) images are generated by subtracting its blurred images as

$$D(x,y,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) = L(x,y,k\sigma) - L(x,y,\sigma). \quad (3)$$

Usually, the local maxima or minima can be found in the edge region, the feature points in spot and corner of the object in DoG images will be mostly selected in DoG images. Through feature points localization, orientation assignments and feature descriptor processing, the final SIFT features are extracted for those feature points.

III. The Proposed CSPR System

A. Overview of the System

The flow diagram of the proposed color-SIFT product recognition (CSPR) system is shown in Fig. 3. This system includes three major stages; training, testing, and classification. As shown in Fig. 2, for the training stage, the CSPR system first needs to collect all images of the target visual things. The color SIFT features of all visual things are extracted and binarized to become color SIFT bit streams. To fasten the search, the bag-of-words [16], [17] with K-means clustering concept is adopted to create a dictionary for classification.

In the testing stage, a picture is shot from a webcam, and the color SIFT features of the test image is extracted and formed the bit stream. The color SIFT features of key-points are submitted to the classification stage, which could be performed in the cloud server. Of course, the trained dictionary is obtained in training phase and could be computed off-line in advance. However, the testing phase and classification phase should be completed in a reasonable short time. The initial classification result will be recognized with its side information from the database after matching the descriptors with the trained dictionary through the SVM procedure [18]. Finally, the SVM result has to be checked if it is reliable. If the SVM score is higher enough, the initial classification will directly output the final decision. If not, we suggest selecting multiple N candidates and performing another verification process by checking moment color information for all matching candidates again. The detail explanations of all key functions will be described in the next subsections.

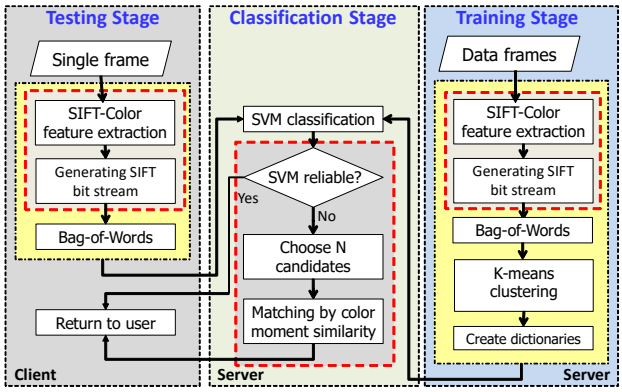

Fig. 3 Flow diagram of the proposed system in three stages

B. Binary Color SIFT Presentation

The color-SIFT descriptor could be achieved by combining the original SIFT feature with color information. To design a better representation of color information, a new color descriptor is proposed. Originally, for transmission and compression purposes, the signs of the features are employed to represent the binarized SIFT descriptor based on the concept of CDVS [11]. As shown in Fig. 4, for each element, the sign of

SIFT will be assigned as "1" for the positive value and "0" for zero value.

Fig. 4 Original binary SIFT descriptor

There are many ways to include the color information in the binary SIFT descriptor. Since the color value is between 0 - 255, it requires 8 bits to represent one color channel, the value "0" can be represented by "00000000", and the value "256" can be represented by "11111111", totally 256 color scales. At the feature point, there are 24 bits for three color channels. To concatenate the SIFT bit stream with 24 bits color information, the direct color-SIFT could be shown as Fig. 5.

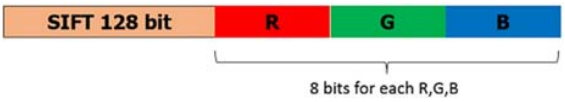

Fig. 5 Direct color-SIFT descriptor with 24-bit RGB format

To achieve the robustness, we suggest the color moments to represent the color information. We select a (2*M*+1) x (2*M*+1) window around the feature point and compute the color moments of each color channel. For *c* (i.e, *R*, *G*, *B*) color channel, three moments of mean E_c^k, standard deviation D_c^k, and skewness S_c^k are respectively computed as:

$$E_c^k = \frac{1}{(2M+1)^2}\sum_{i=-M}^{M}\sum_{j=-M}^{M} p_c^k(x_k+i, y_k+j) \quad E_c^{k*} = \begin{cases} 1 \text{ if } E_c^k > 127 \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

$$D_c^k = \sqrt{\frac{1}{(2M+1)^2}\sum_{i=-M}^{M}\sum_{j=-M}^{M}(p_c^k(x_k+i,y_k+j)-E_c^k)^2} \quad D_c^{k*} = \begin{cases} 1 & \text{if } D_c^k > T_D \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$S_c^k = \sqrt[3]{\frac{1}{(2M+1)^2}\sum_{i=-M}^{M}\sum_{j=-M}^{M}(p_c^k(x_k+i,y_k+j)-E_c^k)^3} \quad S_c^{k*} = \begin{cases} 1 & \text{if } S_c^k > 0 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where P_c^k represents the pixel value at the k^{th} feature point in the c^{th} color channel, (x_k, y_k) means the position of the k^{th} feature point. To achieve the binary color-SIFT, the color moments of feature points also need to be binarized. Three color moments are the robust measures that can be used to differentiate the color similarity in the feature point. After comparing color-SIFT descriptors of each image, some error matching like background feature points can be removed.

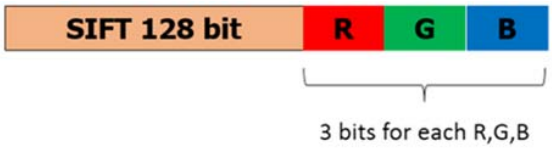

Fig. 6 Moment color-SIFT descriptor with 9-bit RGB format

Finally, as shown in Fig. 6, the moment color-SIFT can be obtained by concatenating 128-bit SIFT with 9-bit color information, which are three moments for three color channels. In summary, the 9-bit RGB color SIFT descriptor as shown in Fig. 6 is chosen in our system. Because it considers surrounding color information around the feature point, it achieves better performance and accurate recognition than the direct 24-bit RGB color-SIFT descriptor after some simulations.

C. Classification by SVM

In this paper, libSVM tools [18] are used to perform the classification. If the prediction score S_P is less than the threshold T_P (=3 in this paper), we will regard it as a trusted prediction. If the prediction score is larger than T_p, the next iteration will be conducted to make the further check. The top S_P classes will be chosen with similar probabilities. Then, the feature points should be matched between test picture and the pictures in the database by K-D tree algorithm [20]. The feature points which are matched by using the minimum Euclidean distance, are smaller than the threshold T_E.

The final step is to discard the unreliable matching as shown in Fig. 7. If the predictive score S_P is larger than the threshold T_P, the nine color bits from color-SIFT between two matching pairs will be compared. If the Euclidean distance of nine color bits of two pairs is greater than T_{EU}, it means that these pairs are not matched in the color domain, and they will be discarded. Finally, if the matching score of any two images is greater than T_m, the recognition result will be returned and show its related object information.

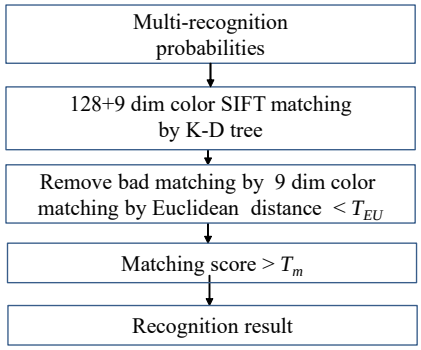

Fig. 7 Flow chart of discarding the unreliable matching

IV. Experimental Results

A. Setup of Experiment Environment

The visual thing recognition systems with different classification methods are implemented by MATLAB 2013 and run in Intel Core i7-4770 CPU 3.40 GHz, 16GB memory and Win7 64-bit computer. All the database and testing pictures are shot by Logitech webcam with the resolution of 1024x768 and are resized to the resolution of 850x640.

B. Discard Unreliable Matches by Color-SIFT

Two image pairs are matched by K-D tree algorithm in gray scale with 128 dim binary SIFT descriptor firstly. If the Euclidean distance of color information is less than threshold

T_{Eu} in a matching pair, this pair is considered as an unreliable matching in color domain. As shown in Fig. 8 (a), the green lines are the matching results by K-D tree algorithm. A clue can be observed that some lines such as the background and color difference regions are matched. This situation is unavoidable because these pairs have similar vectors in gray scale. In Fig. 8 (b), the unreliable matching lines are discarded by comparing their color moments descriptor. The red lines are the discarded matches.

C. Simulations of Objects with Different Color

In case that the test image is blue rose and the data sets have red, blue, and yellow roses, if we only use the original SIFT which is in gray scale to describe feature point, it will be difficult to differentiate which rose is the test image. By using color information around feature points, we can easily recognize that the test image is blue rose. As shown in Table I, if we match images without using color information, the detected result will be yellow roses because the average matches between the test image and yellow roses are highest. But with color information, the highest average matches are blue roses. In this experiment, the original SIFT recognizes it as a yellow rose, while the color-SIFT classifies it as a blue rose. Obviously, the color-SIFT can recognize the color object correctly.

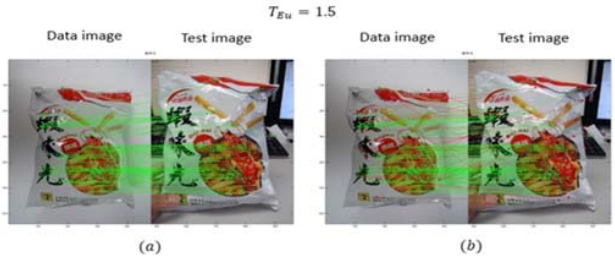

Fig. 8 (a) Matching result by K-D tree (b) Discarding unreliable matches by binary color –SIFT

TABLE I
THE MATCHING LINES OF DIFFERENT COLOR ROSES

Euclidean distance \ Object	Yellow rose	Red rose	Blue rose
Without color	364	347	353
Distance=0	7	1	23
Distance=1	22	18	78
Distance=2	175	168	274

D. Object Recognition

In order to test all possible visual thing recognition systems, the recognize rates of the original SIFT, SURF, and color-SIFT will be presented. 100 visual thing images will be tested by using binary SIFT+BoW [19], binary SURF+BoW [19], LBP+SIFT [21] and proposed system in uses of different color moments. These methods are tested with window size 15 in YUV color model.

In Table II, the simulation results show that the recognition rate in the proposed system with all mean (E), standard deviation (D) and skewness (S) components achieves higher 21% than binary SIFT and 61% than binary SURF. Compared to the LBP-SIFT in [21], the recognition rate is higher than 8%. Although the binary color-SIFT descriptor is binarized in advance, the proposed system still keeps high recognition rate. If we use mean (E) component only or mean (E) and standard deviation components, the color-SIFT will not perform so well.

V. CONCLUSIONS

To overcome the missing of color information in the original SIFT, a color-SIFT descriptor was presented in this paper. The performance of color-SIFT descriptor was better than original SIFT in the situation of color similar objects. We considered color pixel values around feature points, the matching performance was better because that the high qualities of data were not easily to be affected. As experimental results, the recognition rate of the color-SIFT recognition system was higher than original SIFT and SURF, and it still kept high recognition rate although the descriptors were binarized in advance.

TABLE II
RECOGNITION PERFORMANCES OF THE WELL-KNOWN AND THE PROPOSED METHODS

Methods	Binary SIFT + BoW	Binary SURF + BoW	Binary LBP_SIFT+ BoW [21]	Proposed system (E)	Proposed system (E+D)	Proposed system (E+D+S)
Recognition Rate	68%	28%	81%	61%	81%	89%

REFERENCES

[1] D. Zhang, K. H. Yap, S. Subbhuraam, "Mobile Product Recognition with Efferent Bag-of-Phrase Visual Search," *Communications, Control and Signal Processing (ISCCSP)*, pp. 65-68, 2014.
[2] W. Zhang, K. H. Yap, D. J. Zhang, Z. W. Miao, "Feature Weighting in Visual Product Recognition", *Proc. of IEEE International Symposium on Circuits and Systems*, pp.734-737, 2015.
[3] S.-M. Huang and J.-F. Yang, "Improved Principal Component Regression for Face Recognition under Illumination Variations", *IEEE Signal Processing Letter*, vol. 19, no. 4, pp. 179-182, April 2012.
[4] C.-Y. Su and J.-F. Yang, "Histogram of Gradient Phases: A New Local Descriptor for Face Recognition", *IET Computer Vision*, vol. 8, no.6, pp.556-567, December 2014.
[5] D. Lowe, "Object Recognition from Local Scale-invariant Features", *Proceedings of the International Conference on Computer Vision*," pp. 1150–1157, 1999.
[6] D. Lowe, "Distinctive Image Features from Scale-invariant Key-points", *International Journal of Computer Vision*, vol. 60, no. 2, pp.91–110, 2004.
[7] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speed-Up Robust Features", in *Computer Vision and Image Understanding* (CYIU), vol. 110, no. 3, pp. 349-359, 2008.
[8] J. Matas, O. Chum, M. Urban, T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions", *Proc. of British Machine Vision Conference*, pp. 384-396, 2002.
[9] N. Dalal, B. Triggs, "Histogram of Oriented Gradients for Human Detection", *Proc. of IEEE Conference on Computer Vision and pattern Recognition (CVPR'05)*, vol. 1, pp. 886-893, 2005.
[10] C. Harris, M. Stephens, "A Combined Corner and Edge Detector", P*roc, of the 4-th Alvey Vision Conference*, pp. 147-151, 1988.
[11] R. O. Duda, P. E. Hart, "Use of the Hough Transform Translation to Detect Lines and Curves in Pictures," *Comm. ACM*, vol. 15, pp. 11-15, 1972.

[12] J. Shi, C. Tomasi, "Good Feature to Track", *Prof. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR94)*, pp. 593-600, 1994.
[13] P. M. Panchal, S. R. Panchal, S. K. Shah, "A Comparison of SIFT and SURF", in *International Journal of Innovative Research in computer and Communication Engineering* vol. 1, no. 2, pp. 323-327, 2013.
[14] L, Bo, T. Whangbo, "A SIFT-Color Moments Descriptor for Object Recognition", *Proc. of International Conference on IT Convergence and Security (ICITCS)*, pp. 1-3, 2014.
[15] L. Y. Duan, F. Gao, J, Chen, J. Lin, T, Huang, "Compact Descriptor for Mobile Visual Search and MPEG CDVS Standardization", *Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 885-888, 2013.
[16] S. Josef, Z. Andrew, "Efficient Visual Search of Videos Cast as Text Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp.591-605, 2009.
[17] D. Nister and H. Stewenius, "Scalable Recognition with a vocabulary tree" in *Proc. of IEEE Conference CVPR*, pp. 2161-2168, 2006.
[18] C. C. Chung, C. J. Lin, "LibSVM" http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[19] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-features Image Classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 490–503, 2006.
[20] C. Sharat. "Introduction to kd-trees", University of Maryland Department of Computer Science.
[21] L. Kabbai, A. Azaza, M. Abdellaoui, A, Douik "Image Matching Based on LBP and SIFT Descriptor", *Prof. of IEEE Conference on Systems, Signals & Devices (SSD)*, pp.1-6, 2015.