Augmented Visual Phrase in Mobile Product Recognition

Wen Zhang, Anu Susan Skaria, Dipu Manandhar, Kim-Hui Yap and Zhenwei Miao
School of Electrical and Electronics Engineering
Nanyang Technological University,
Singapore 639798.
Email: {WZHANG017, ANUSUSAN001, DIPU002, EKHYAP, ZWMIAO}@ntu.edu.sg

***Abstract*— With the rapid advancement in mobile device technologies and connectivity, the use of mobile devices for visual object recognition is emerging as an application with great commercialization potentials. However, query images captured by mobile devices often suffer from various conditions such as illumination, scale, and viewpoint changes. To handle these, several detectors and descriptors have been proposed. However, recognition remains a challenge under strong photometric or geometric variation. In view of this, we propose a new Augmented Visual Phrase (AVP) framework that addresses this issue by using augmented features from transformed images. We propose a recognition framework based on the Bag of Phrase (BoP) structure which in turn is built on the Bag of Words (BoW) model. The proposed method can provide better performance by incorporating the spatial relationship of visual elements detected by keypoint detectors. To further eliminate spurious matches of visual phrases, Geometric Verification (GV) is applied to the top-ranked images. Experimental results show that the proposed AVP method outperforms the current BoP method by 9% in recognition rate.**

Keywords- Bag of Phrase, Augmented Visual Phrases (AVPs), Photometric and Geometric Distortion

I. INTRODUCTION

The tremendous development in technology has brought us mobile devices like smart phones and tablets that possess high computational power with strong imaging capability. Coupled with wireless connectivity, these devices are a platform ready for mobile commerce. Amongst these, mobile visual product recognition and recommendation is emerging as an appealing application, where products are visually recognized and customers are offered recommendations that enhance online purchase experience.

In practise, the query images of a product taken from handheld mobile devices are accquired under diverse imaging environments. They suffer from various conditions such as lighting, scale, perspective changes, etc. These tend to create query images that differ significantly from the reference images in the database. This poses a great challenge in visual recognition.

One of the problems in image retrieval and recognition systems is keypoint drop-out or failure in detecting potential interest points. A good detector should have good repeatability and consistency. Many detectors have been proposed to improve these charateristics [11], [14-15]. However, even a widely accepted keypoint detector such as scale-invariant feature transform (SIFT) detector [11] has limited tolerance to geometric and photometric transformation. Potential keypoints may be undetected due to these variations which may cause failure in recognition. This situation can be visualized in Figure 1 which shows images of an object under different illuminations. There are many additional new keypoints detected (shown in red) in the photometric-transformed images (b) and (c). These new keypoints are not available in the orignal image (a), and it can impact on the recognition results. In addition, even for the common detected keypoints (shown in green), their descriptors are noisy and may result in incorrect quantization of visual words. In short, SIFT descriptors has shown good performance under different conditions of images [13]. However, it still cannot handle photometric and certain geometric variation well enough.

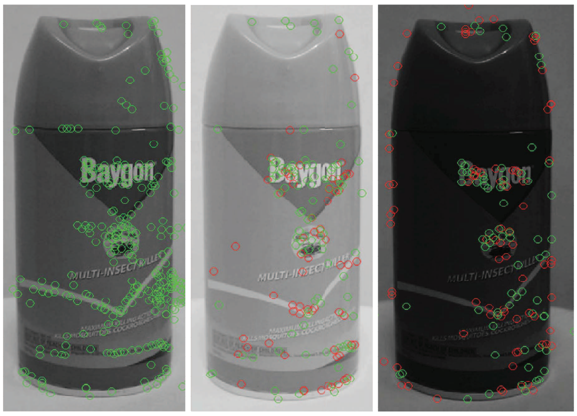

Figure 1. Images under different illumination conditions with keypoints. a) Original image and keypoints in green color, b) and c) two gamma transformed images with additional keypoints in red color.

BoW model [1] has been widely used in image retrieval and recognition [1-4], where visual vocabulary is constructed using clustering algorithms. The raw descriptors extracted from image are quantized to discrete visual words. Finally, an image is represented as a vector of visual words which are directly used for checking image similarity. However, even a finely grained visual word in BoW method has limited descriptive and discriminative ability [5]. This is because this approach ignores the spatial information of visual words. Spatial

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.

verification is only performed on high-ranked images retrieved by initial filtering using the BoW model.

Several methods [5-10] have been proposed to incorporate certain spatial information to form BoP model, built upon the BoW structure. Visual phrases are more discriminative and descriptive than visual words. A visual phrase match is a strong evidence of presence of a similar patch in the images. However, the performance of the BoP and BoW methods still suffer from the issues of photometric and certain geometric distortions as discussed earlier.

In the BoP approaches, visual words co-occuring within certain spatial region are selected as discriminative visual phrases [5], [7] and [10]. An image is represented as a histogram of these selected visual phrases. As these methods only consider the original database images in visual phrase generation and selection, they are unable to handle query images which may suffer from illumination and geometric distortions. This is because the detectors has limited repeatability under diverse imaging conditions. This, in turn, leads to missing potential visual phrase which can be significantly impact the recognition performance for a query image.

To handle these issues, we propose a framework that uses Augmented Visual Phrases (AVP) in the BoP model. We select AVPs from a pool of augmented features constructed by merging all the discriminative keypoints and their corresponding descriptors from the original as well as the transformed images. These selected visual phrase are meaningful as they incorporate features carrying more diverse information. The improvement in recognition rate shows that the AVP method can alleviate the challenge caused by the above-mentioned distortions.

The rest of the paper is organized as follows. Section II introduces the overview of our proposed AVP algorithm. Section III describes the transformation, generation and selection of augmented features for an image. Section IV presents the AVP candidate generation and selection. The method of efficient 2D indexing is explained in Section V followed by experiments, and discussion and conclusion in Section VI and VII respectively.

II. PROSOSED AVP FRAMEWORK

The flowchart of the proposed AVP method for image recognition is shown in Figure 2. It consists of two main phases: 1) offline phase and 2) online phase.

Offline phase is the training phase based on the augmented features extracted from all the reference images in the database. In order to train the system that can handle variations caused by illumination, perspective change etc., we need to augment features from such distorted images. To do so, an original image is subjected to different artificial photometric transformations (illumination changes) and geometric transformations. This is similar to that in [12] to generate the augmented pools of SIFT keypoints. This step will be explained in greater details in Section III. The augmented features from all database images are used to train a scalable vocabulary tree (SVT) using hierarchical K-means [2] to generate a visual vocabulary of N leaf nodes. With the trained SVT, the augmented features are quantized to a set of visual words. Note that the generated visual words carry information of the original as well as the transformed images.

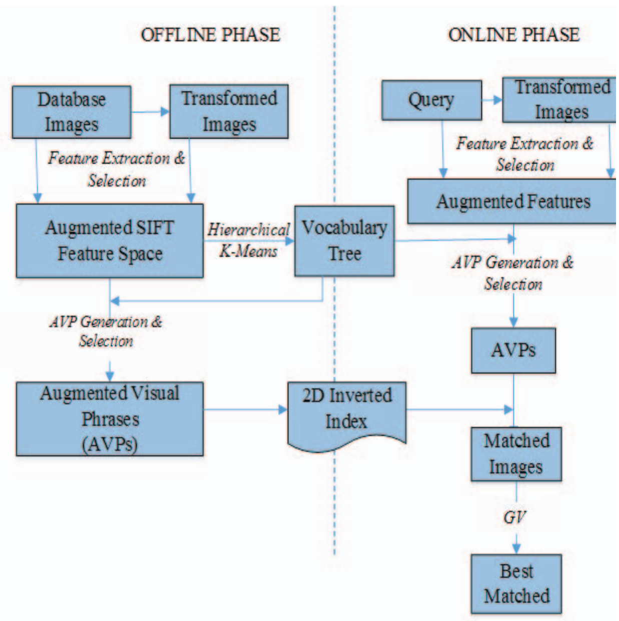

Figure 2. Proposed AVP method for visual recognition

To capture pairwise spatial relation in the visual words, two visual words from the augmented features are paired up to form second-order visual phrase. These visual phrases integrate two properties that can enhance the recognition performance.

1) Spatial information to make features more discriminative;
2) Augmented features to handle potential distortion.

Unreliable visual phrase are filtered out through AVP selection and generation based on frequency analysis of visual phrases. Further discussion on AVP selection will be given in Section IV. Eventually, database is indexed by 2D inverted index based on selected AVPs.

In the online phase, augmented features from the query image are extracted and visual phrases are constructed using the same method as for the reference image. Online matching with the database images is done using an efficient 2D inverted index. However, as there exist some spurious AVPs, to ensure a confident matching, RANSAC-based geometric verification (GV) is used on the top-ranked retrieved image to achieve good geometric consistency of the visual phrases. Subsequently, the best matched image is retrieved and the product is recognized.

III. Augmented Features

Our aim is to handle the issue that query image may experience in different imaging conditions from the database images using AVPs. To address this, we artificially generate images under different simulated conditions. We apply two types of transformations, namely, photometric and geometric transformations.

A. Phtometric Transformation

For photometric transformation, we used a non-linear effect popularly known as gamma correction to adjust the image contrast.

$$G(u) = u^{\frac{1}{\gamma}} \qquad (1)$$

where u is a normalized pixel with u □ [0,1] and γ is the parameter which is varied to generate brighter or darker images. Different values of γ are chosen to create transformed images with different illuminations. We have chosen a range a values for the gamma parameter as follows, γ = {2.1^{-2}, 2.1^{-1}, 2.1^{0}, 2.1^{1}, 2.1^{2}}. Here, we have four transformed images. The features from these images can help to characterize images that experience different illuminations.

B. Geometric Transformation

Further, we have also included features from geometric transformed images. To simulate different geometric transformations, new images are generated by affine transformation using a 2×2 transformation matrix T. A new location of each pixel in the 2D coordinate is obtained by multiplying the current coordinate with the matrix T. Different parameter settings for the matrix T define different transformations. Four rotated and four scaled versions of image are generated to handle the perspective and scale variations. Specifically, we used rotation angle θ = {10º, 20º, 30º, 40º} and scale factor s = {1.2, 1.4, 1.6, 1.8}.

After combining both the photometric and geometric transformations, we have t = 12 transformed images for an original database image. However, not all of these features from the transformed images are discriminative. Thus, we only select those discriminative features that are compatible with features in the original image. The compatibility criteria are: 1) the keypoint in the transformed image should correspond to a neighboring keypoint that is within a radius of 2 pixels from the original image, and 2) the corresponding descriptors must have similarity greater than 0.94. Similarity is given by $sim(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}*\boldsymbol{Y}/(|\boldsymbol{X}|*|\boldsymbol{Y}|)$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the two descriptors.

A database image is now augmented into a set of transformed images: {$I(n)$} where n □ {0,1,2,···,t}, where $I(0)$ represents the original image. Let $D(n)$ be the SIFT features set extracted from image $I(n)$. We select discriminative features $\Psi(n)$, which is a subset of $D(n)$ that satisfy the above compatibility criteria. All these discriminative features are merged to form a pool of augmented features for an image, namely, $A = Unique(\{\Psi\}_{n=0}^{t})$. The augmented features A from all database images are used to construct a SVT using hierarchical K-means clustering algorithm [2] to generate N visual words.

Algorithm 1: AVP Generation Algorithm

```
Input:  Database Images
Output: Augmented Visual Phrases (AVPs)

1) For all images in the database,
       Generate transformed images: $I(n)$, where $n$ □ {$0,1,2,\cdots,t$}
           For $n = 0 : t$
               Extract all SIFT features sets $d(n)$
               Select $\Psi(n)$ from $d(n)$ that satisfy compatibility criteria
           End
       Augmented Feature $A = Unique(\{\Psi\}_{n=0}^{t})$
   End

2) SVT training in Augmented Features Space $A$
       Choose Branch-factor $K = 10$, depth $L = 5$
   Output: SVT with N visual words

3) AVP Selection and Generation
   For all images,
       Quantize $A$ into visual words $W = \{w_i\}$, where $i$ □ {$1, 2, \cdots, N$}
       If $w_j$ is in the neighboring circular region of $w_i$ with a radius
       of $r_i = scale_i \cdot \lambda$
           Select $AVP$, $\Omega = \{(w_i, w_j) \mid 1<i,j<N\}$
       End
   End
   Set $Score(i, C) = count(i, C) / \ln(count(i, Z)$
   Discard visual phrase having $Score < Thres$.

4) Return AVPs
```

IV. AVP Candidate Generation and Selection

First of all, the pool of augmented features for all images is quantized using SVT to form augmented visual words which are used for generation of augmented visual phrases (AVPs).

The AVP generation and selection procedure using the proposed method is summarized in Algorithm 1.

A set of visual words that are co-occurring within certain spatial region forms a visual phrase. Visual phrases of higher order are more discriminative; but very sparse and complex for realization. Thus for simplicity, only second order visual phrase composed by two visual words are considered. A second order visual phrase is an ordered pair of two visual words: Ω = {(w_i,w_j) |1<i,j<N}.Thus, augmented visual phrase for an image are such ordered pairs from augmented visual words satisfying the following spatial relation.

The criteria for AVPs is that they should be co-occurring in neighboring circular region defined by radius $r = (scale \cdot \lambda)$, where $scale$ is the scale of the keypoint detected using SIFT detector [11] . λ is chosen to be 5 experimentally.

The initial candidates for visual phrase are huge in number and may contain many unstable and non-representative visual phrases. To select the discriminative visual phrase, we have to discard weak and unreliable pairs that have lower frequency. To prioritize discriminative visual phrase for a particular category, a score based on TF-IDF scheme in (2) is adopted [5]. Only those AVPs that have score greater than a threshold are retained. We selected this threshold to be 0.8 experimentally.

$$Score(i,C) = count(i,C)/\ln(count(i,Z)) \qquad (2)$$

where *count*(*i,C*)and *count*(*i,Z*)are number of visual phrase i in category C and all database images Z, respectively. The score value is a measure for discriminative ability of the visual phrase.

V. EFFICIENT 2D INDEXING

For set of second order visual phrase Ω, the number of visual phrase in Ω is N^2, where N be the number of visual words in SVT. This is a very large number. Indexing such a large quantities requires huge memory. As the selected visual phrases in Ω are much smaller than N^2, to efficiently index AVPs, we follow the method proposed in [10] to perform 2D inverted indexing.

A visual phrase of second order; $\Omega = \{w_i, w_j \mid 1<i,j<N\}$ is indexed using a N×N dimension matrix, where w_i and w_j are co-occurring visual words from the augmented features. To take the advantage of the sparsity of matrix, we create 1D inverted index of first dimension w_i; where i □ {1, 2,···,N} and list w_j in another dimension while discarding all the null entries. Since the symmetric visual phrase ($w_{i,j}$ & $w_{j,i}$) are repeated visual phrases. Thus, to further reduce storage, we only keep those $\Omega=\{w_i, w_j\}$ satisfying $i < j$. Subsequently, an image is represented as a histogram of these visual phrases Ω.

To find the image similarity, histogram intersection is used as given by (3):

$$\mathrm{I(D,Q)} = \sum_{n=1}^{M} \min\{d^n, q^n\} \qquad (3)$$

where d^n and q^n are the n^{th} components of the histogram of the database image D and the query image Q, respectively.

Next, to eliminate spurious matches due to background clutters and mismatched AVP pairs in the augmented space, the top 50 ranked image are subjected to geometric verification based on RANSAC algorithm. To test the performance of the proposed method, we use recognition rate as the performance metric, which considers the result to be correctly recognized if and only if the first- ranked image belongs to the true ground truth category.

VI. EXPERIMENTAL EVALUATION

We build a commercial product database which contains 3882 reference and 333 test images from 41 categories to evaluate the proposed method. All the reference images are captured by digital cameras under good imaging condition, whereas all the test image are captured using mobile phones. These test images are acquired under different conditions, e.g. different sizes and viewpoints, illumination and background clutters, etc. Some selected sample reference and test images are shown in Figure 3.

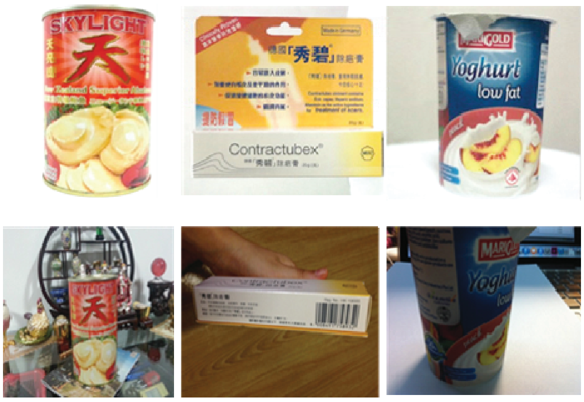

Figure 3. Sample images from commercial product database. First and second rows show the reference and test images respectively.

In the experiments, SIFT is used to detect the keypoints. Augmented features are formed by selecting discriminative features from the reference and transformed images. These pool of augmented features are then used to construct SVT with a branch factor of 10 and depth of 5, giving rise to N=100000 leaf nodes. Subsequently, an image is represented as a histogram of selected visual phrase using AVP method, and indexed using the method described in Section V. To handle spurious matches and obtain good match, we apply GV for top 50 ranked images.

The performance comparison of the proposed AVP method with respect to the developed BoP method in [10] and the implemented SVT method in [2] is shown in Table 1.

The proposed method achieves a recognition rate of 93.1% which outperform the SVT method by 12.3%. This improvement is benefited from the advantage of the AVP method that it incorporates the spatial information whereas SVT discards the spatial features. Further, the proposed method outperforms the efficient BoP by 9.1%. Even though the efficient BoP method in [10] uses spatial information, it cannot handle query images with high photometric and strong geometric distortions. Overall, the proposed method can address these issues and achieves a good performance.

TABLE I. PERFORMANCE EVALUATION OF THE PROPOSED METHOD WITH THE BOP [10] AND SVT METHODS

Method	*BoW using SVT [2]*	*BoP* [10]	*Proposed AVP Technique*
Recognition Rate (%)	82.9%	85.3%	93.1%

A case study of our proposed method against the BoW using SVT [2] and the efficient BoP method [10] is shown in

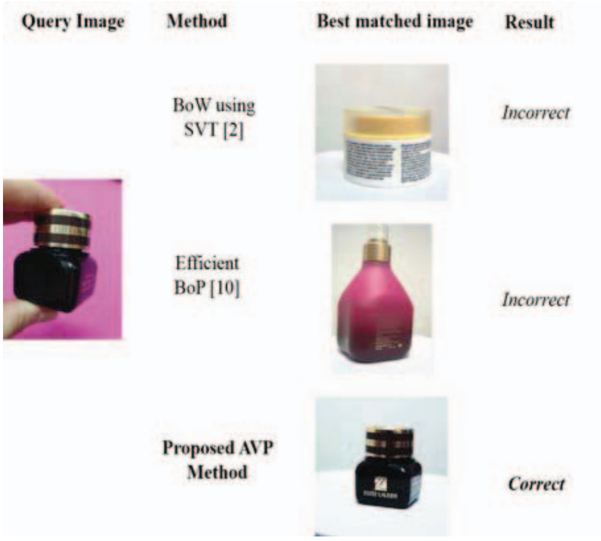

Figure 4. Comparison of the proposed method with the BoW using SVT [2] and the efficient BoP [10].

Figure 4. From the figure, it can be seen that although the query image is taken under different perspective and illumination condition as compared to the reference database image, the proposed method is able to recognize the product correctly, whereas the other methods fail.

VII. Conclusion

In this paper, we proposed the AVP algorithm that can address the issue of photometric and geometric distortions in the query images. We generated augmented features using transformed images. Using spatial relationship to make the feature more discriminative and using augmented features to handle variability in query images, the proposed method is able to perform recognition of consumer product with different illumination and perspective changes. Experimental results shows the proposed method can achieve a good recognition rate of 93.1%, and outperform other current methods.

References

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, 2003, pp. 1470-1477.

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

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, pp. 1-8.

[4] K.-H. Yap, T. Chen, Z. Li, and K. Wu, "A comparative study of mobile-based landmark recognition techniques,"Intelligent Systems, IEEE, vol. 25, pp. 48-57, 2010.

[5] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li, "Descriptive visual words and visual phrases for image applications," in Proceedings of the 17th ACM international conference on Multimedia, 2009, pp. 75-84.

[6] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, pp. 1-8.

[7] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative bag-of- visual phrase learning for landmark recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012, pp. 893-896.

[8] Y. Jiang, J. Meng, and J. Yuan, "Grid-based local feature bundling for efficient object search and localization," in Image Processing (ICIP), 2011 18th IEEE International Conference on, 2011, pp. 113-116.

[9] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 809–816, 2011.

[10] D. Zhang, K.-H. Yap, and S. Subbhuraam. "Mobile product recognition with efficient Bag-of-Phrase visual search." *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*. IEEE, 2014.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, pp. 91-110, 2004.

[12] W. Zhang, K.-H Yap, D. Zhang, and Z. W. Miao, "Feature Weighting in Visual Product Recognition." *IEEE International Symposium on Circuits and Systems, Lisbon*, Portugal, pp. 1-4 2015.

[13] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, in: CVPR, vol. 2, June 2003, pp. 257–263.

[14] 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.

[15] Z. W. Miao, X. D. Jiang and K.-H Yap, "Contrast Invariant Interest Point Detection by Zero-Norm LoG Filter," IEEE Transactions on Image Processing, DOI: 10.1109/TIP.2015.2470598