Where's the Weet-Bix?

Yuhang Zhang, Lei Wang, Richard Hartley, and Hongdong Li*

Research School of Information Sciences and Engineering
Australian National University

Abstract. This paper proposes a new retrieval problem and conducts the initial study. This problem aims at finding the location of an item in a supermarket by means of visual retrieval. It is modelled as object-based retrieval and approached using the local invariant features. Two existing retrieval methods are investigated and their similarity measures are modified to better fit this new problem. More importantly, through the study this new retrieval problem proves itself to be a challenging task. An instant application of it is to help the customer find what they want without physically wandering around the shelves but a wide range of potential applications could be expected.

1 Introduction

Given the query image of an object, object retrieval requires to find the same objects from a collection of images [1,2]. In this paper, we propose a new object retrieval problem, in which the object to retrieve only occupies a small part of a database image and might have multiple copies there. An instant application is: suppose a customer needs a certain brand of biscuit and he has a sample or its image. Through our object retrieval, the customer will be informed the shelf where this biscuit lies. This problem is more challenging than those posed in [1,2] due to the following issues:

1. The query image is a close-up view of the object to find, whereas each database image contains dozens of objects that are small in size and different in brands and manufacturers, as shown in Figure 1. As a result, the object to retrieve is presented in very different scales between the query and database images. Worse is that in a database image all the objects other than the queried become background clutter.
2. In each database image, there are often multiple copies of an object. In that case the same or similar local invariant features may be extracted from different locations, which makes some widely used geometrical constraints not suitable anymore. For example, some features points found in a database image may not be close to each other, even if their matches are close on the query image;

* The last two authors are also affiliated with NICTA, a research institute funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

- 3. To attract the attention of customers, the appearances of objects in supermarket are often full of striking signs and patterns, leading to a very rich set of local features. This challenges the discriminating ability of the descriptors for different local features. Moreover, a large number of objects have round or non-rigid shapes and glisten are very common. These situations further increase the difficulty of this object retrieval task.
- 4. Products are often produced in series, for instance, “BIO” brand includes the detergent for brighter color and the softener. In this case, the appearances of objects have only slight difference. However, they are completely different for customers. This fact is reflected in the ground truth defined for this retrieval problem, as shown in Figure 2.

In this paper, the image database is created and its characteristics are discussed. From each image, local invariant regions are detected and described by the SIFT feature [3]. Upon all of these features, a visual vocabulary is constructed. Each image is then represented by a histogram vector showing the occurrence frequency of each visual word. The similarity of two images is measured by comparing the corresponding histogram vectors. To address the issues aforementioned, a loose geometrical constraint is applied, which partitions each database image into multiple overlapped sub-images to mitigate the affect of background clutter. Taking the advantage of the existence of multiple copies,

Fig. 1. Top row: query images; Bottom row: database images

Fig. 2. Definition of ground truth (Must be identical rather than similar only)

new weighting scheme is designed to assign high weight to the visual words intensively presented in a small number of database images. An inner-product based similarity measure is developed, which encourages the “co-presence” of same visual words in query and database images but does not penalize the other cases like the “co-presence” of different visual words. Experimental study is conducted to compare our approach with the existing methods from [1,2] on this new retrieval problem, demonstrating its superior performance.

2 Related Work

This paper employs local invariant features to represent the visual content of an image. The interest point detecting approach Harris-Affine detector proposed by Krystian Mikolajczyk and Cordelia Schmid in [4] provides interested point for reliable matching for images even with significant perspective deformations. In later discussion in [5], it is also demonstrated to be more efficient than most other region detectors when dealing with occlusion and clutter. The image feature generation approach proposed by David G. Lowe in [3] gives SIFT features which are largely invariant to changes in scale, illumination, and local affine distortions. These local-invariant features are of intermediate complexity, which means they are not only distinctive enough to determine matches in big feature database but also sufficient enough to bear clutter and occlusion.

There has been considerable progress in developing real-world objects recognition systems with local invariant features in recent work [1,2,6,7]. Among those work Visual vocabulary [1] has been proved to be a distinctive indexing artifice in local invariant feature based object retrieval. To build a visual vocabulary, local feature descriptors are quantized into clusters according to their similarity. Then when a new local feature comes, all the similar features can be found by assigning the new feature’s descriptor into its nearest cluster.

In [1] retrieval of all occurrences of an object outlined by user from a movie is carried out. After extracting local invariant features from each frame, two visual vocabularies of different types of local invariant features are built up. As the complement of visual vocabulary, term frequency-inverse document frequency (tf-idf) weighting standard, L_2 -norm similarity measure, and spatial consistency check are employed to improve the retrieval performance, showing excellent retrieval performance. In [2], visual vocabulary is extended into a hierarchical vocabulary tree and a scheme which can quickly retrieve the same CD-covers from a large database of music CDs is proposed.

In [2], all the query and database images contain only a single object, namely CD cover. Their work tries to find out whether a query and a database image have the same content. However in our work, although a query image contains single object, database image always contains dozens of objects of different classes. What we try to find out is whether a database image contains the single object on a query image. In [1] the problem seems to be the same with us, but their query images are cropped from database images, which actually results in finding out where an image patch comes from. Moreover, a new issue we need to deal

with is the potential multiple appearances of query object in single database image.

3 Creation of the Database

Our database named *WebMarket* contains 3,153 images which were taken in a supermarket named Coles in Jan 2007. This supermarket has eighteen 30-meter-long shelves, each of which approximately has six levels. Ten shelves are captured in this database. Starting from one end of the first shelves, the photographing was carried out roughly following the order of the shelves. Two adjacent images have some overlap (less than one third of image size) to ensure that no part of the shelf is missed. Three digital cameras were used and all the images are saved in Jpeg format. The resolution of the images is either $2,272 \times 1,704$ or $2,592 \times 1,944$ pixels. Each image generally covers an area of about 1.5 meter in height and 2 meters in width on shelves, imaging all the objects within three or four shelf levels in range. The size of each single object in an image is usually small. During photographing, no restriction on the view point and distance was imposed, although most of images are frontal views. No special illumination is used. To build the query set, about 200 different objects were randomly selected and put on the ground, and captured one by one. For each query three images are taken from different view angles or distances. For the same object, there is a large difference in its scale between the query and database images.

This image database named *WebMarket* will soon be published on web.

4 Our Approach

4.1 Image Representation

Local invariant features have shown excellent performance in representing images under a certain degree of change on scale, view angle, and illuminance and partial occlusion. In this work, a Harris-Affine interest region detector [4] is applied to each image and the SIFT descriptor [3] is used to describe the detected interest regions. The binaries are downloaded from the Visual Geometry Group, Univ. of Oxford, and the threshold parameter is set to 20,000 to control the number of detected regions. For each database image, about 2,000 SIFT features are extracted, leading to 6.8 million SIFT features in total. When dealing with query images, to handle the large scale difference between the query and database images, the object is manually segmented out (an automatic segmentation algorithm can also be used) and downsized to a 300×300 image to partially alleviate the scale problem. The local features are then extracted from it.

4.2 Construction of the Visual Vocabulary

To ensure the quality of the visual vocabulary, it is built upon all of the 6.8 million SIFT features rather than a randomly selected subset. Hierarchical k -means clustering [8] is applied. Step by step, all of the features in database are

clustered into 1,000, 20,000, and 200,000 clusters, leading to a visual vocabulary containing 200,000 visual words. The hierarchical clustering is stopped at 200,000 because it gives reasonably good retrieval performance on our problem. A larger-sized vocabulary may be used. However, it will run a risk that the features which have been sufficiently similar (up to some noise level) are separated into different clusters and this will adversely affect the retrieval.

4.3 A Loose Geometry Constraint

In our problem, the local features are much richer in a database image than in a query image. As a result, a high matching score can be easily obtained between a query image and most database images, leading to very poor retrieval performance. Spatial consistency check in [1,3] cannot be used here, because there are often multiple copies of identical object in a database image and the identical local features could be extracted from different places.

This paper imposes a loose geometry constraint. It evenly partitions a database image into 25 sub-images, each of which is one ninth of the original one. A sub-image is large enough to contain the object to retrieve but has much less background clutter.¹ Two neighboring sub-images have the half area overlapped to reduce the risk of separating one object into two sub-images. Totally, about 78,000 sub-images are obtained. For a query image, the match of a sub-image means the match of the corresponding database image. After partition, the computational load of retrieval remains the same.

4.4 Similarity Measure

A similarity measure defines the visual similarity between a query and the sub-images. Those with higher measure values are selected as the retrieval result.

Let $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{in}]^\top$ denote the i -th sub-image, where x_{ij} ($1 \leq i \leq m, 1 \leq j \leq n$) is the number of occurrences of the j -th visual word in this image, m the total number of sub-images, and n the total number of visual words. Similarly, the i -th query image is represented as $\mathbf{q} = [q_1, q_2, \dots, q_n]^\top$. In our proposed similarity measure, $\mathbf{q}$ and $\mathbf{x}_i$ are mapped respectively to $\tilde{\mathbf{q}} = [\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_n]^\top$ and $\tilde{\mathbf{x}}_i = [\tilde{x}_{i1}, \tilde{x}_{i2}, \dots, \tilde{x}_{in}]^\top$, where

$$\tilde{q}_j = \begin{cases} 0, & \text{if } q_j = 0 \\ 1, & \text{if } q_j > 0 \end{cases}; \quad \tilde{x}_{ij} = \begin{cases} 0, & \text{if } x_{ij} = 0 \\ d_j, & \text{if } x_{ij} > 0 \end{cases} \quad (1)$$

where

$$d_j = \frac{\sum_{k=1}^m x_{kj}}{\sum_{k=1}^m \text{sign}(x_{kj})} \log \frac{m}{\sum_{k=1}^m \text{sign}(x_{kj})}; \quad (2)$$

$$\text{sign}(x_{ij}) = \begin{cases} 0, & \text{if } x_{ij} = 0 \\ 1, & \text{if } x_{ij} > 0 \end{cases}. \quad (3)$$

¹ An ideal way to implement this constraint may be to ensure that each sub-image exactly contains one object. However, this is no less difficult than the object retrieval problem itself.

The similarity measure is defined as the inner product between $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{x}}_i$,

$$S_1(\mathbf{q}, \mathbf{x}_i) = \langle \tilde{\mathbf{q}}, \tilde{\mathbf{x}}_i \rangle \quad (4)$$

Equation (2) shows the weighting scheme designed for the problem we have. It is a product of two terms: $\sum_{k=1}^m x_{kj} / \sum_{k=1}^m \text{sign}(x_{kj})$ and $\log(m / \sum_{k=1}^m \text{sign}(x_{kj}))$. Here, $\sum_{k=1}^m x_{kj}$ represents the total number of occurrences of the j -th visual word through the database, and $\sum_{k=1}^m \text{sign}(x_{kj})$ represents the number of images containing at least one j -th visual word through the database. If a visual word appears in certain image(s) with high repetition, then it is probably a stable feature that can be extracted from different copies of a same object. On the other hand, if a feature appears in different images but only once in each, it is more likely to be a noise. These two cases are weighted via the first term. Moreover, for a very popular feature that can be extracted from many images, no matter if stable or not, its importance will be down scaled via the second term since it is not discriminating. In addition, the similarity measure computes the inner product of two unnormalized vectors. By doing so, the ‘‘co-presence’’ of the same visual words is rewarded between two compared images. Meanwhile, the case that a visual word only appears in one of them will not be penalized.

We also tried the similarity measure which considers the exact number, x_{ij} , of a visual word appeared in a sub-image. It is defined as

$$\tilde{q}_j = \begin{cases} 0, & \text{if } q_j = 0 \\ 1, & \text{if } q_j > 0 \end{cases} ; \quad \tilde{x}'_{ij} = \begin{cases} 0, & \text{if } x_{ij} = 0 \\ x_{ij}d_j, & \text{if } x_{ij} > 0 \end{cases} \quad (5)$$

$$S_2(\mathbf{q}, \mathbf{x}_i) = \langle \tilde{\mathbf{q}}, \tilde{\mathbf{x}}'_i \rangle. \quad (6)$$

Theoretically, similarity measure denoted by Equation (6) awards the database images who have more copies of the visual words that query image contains, which indicates a higher possibility of true matching rather than noise. However, in this manner those database images who only has single copy of queried object may loose the game against those who don't contain queried object but possess certain amount of similar noise. On the other hand, similarity measure denoted by Equation (2) focus more on how many types of visual word are matched, which is expected to perform better in occasion of that few copies of queried object are presented in a database image.

Another two similarity measure have been proposed in [1,2]. In [1], each query or database image i is mapped to a n -dimensional vector $\mathbf{V}_i = [t_{i1}, t_{i2}, \dots, t_{in}]^T$, where

$$t_{ij} = \frac{x_{ij}}{\sum_{k=1}^n x_{ik}} \log \frac{m}{\sum_{k=1}^m x_{kj}}. \quad (7)$$

Then the similarity between each pair of images is measured by

$$S_3(\mathbf{V}_q, \mathbf{V}_{x_i}) = \langle \frac{\mathbf{V}_q}{\|\mathbf{V}_q\|}, \frac{\mathbf{V}_{x_i}}{\|\mathbf{V}_{x_i}\|} \rangle. \quad (8)$$

In [2] each query or database image i is also mapped to a n -dimensional vector $\tilde{\mathbf{V}}_i = [\tilde{t}_{i1}, \tilde{t}_{i2}, \dots, \tilde{t}_{in}]^\top$, where

$$\tilde{t}_{ij} = x_{ij} \log \frac{m}{\sum_{k=1}^m \text{sign}(x_{kj})}. \quad (9)$$

Then the similarity between each pair of images is measured by

$$S_4(\tilde{\mathbf{V}}_q, \tilde{\mathbf{V}}_{x_i}) = \left\| \frac{\tilde{\mathbf{V}}_q}{\|\tilde{\mathbf{V}}_q\|} - \frac{\tilde{\mathbf{V}}_{x_i}}{\|\tilde{\mathbf{V}}_{x_i}\|} \right\|. \quad (10)$$

All the four similarity measure methods will be compared in the experiments.

5 Experimental Result

This experiment investigates the performance of the proposed similarity measures S_1 and S_2 , as well as the previous similarity measure proposed in [1,2] which are denoted by S_3 and S_4 respectively on this new retrieval task. All of the 3,153 database images are used. Thirty retrievals are conducted and the average retrieval performance is reported. The 30 query images are randomly sampled from the query set containing 600 images. The ground truth of each of them is created by manually checking the database images one by one. On average, for a query image, there are merely 6.63 database images which are the true matches. The retrieval rank of a database image is determined as the highest rank of the sub-images cropped from it. Each image is represented as a 200,000-dimensional vector. Although in very high dimension, this vector is quite sparse. The total number of non-zero items is no more than the total number of local features extracted from this image, about 2,000 in our case. Taking the advantage of the sparsity allows us to efficiently evaluate the similarity of two images. The retrieval performance is measured by the *Precision* and *Recall* widely used in information retrieval. They are defined as

$$\begin{aligned} Precision &= \frac{\#positive_retrieved}{\#retrieved}, \\ Recall &= \frac{\#positive_retrieved}{\#total_positive} \end{aligned} \quad (11)$$

where $\#positive_retrieved$ represents the number of correctly retrieved images, $\#retrieved$ represents the number of retrieved images, and $\#total_positive$ represents the number of the ground truth in the database for a query. Also, the average normalized rank of true matching images in [1] is also used, which is defined as

$$\text{avg_rank} = \frac{1}{m \cdot m_{pos}} \left(\sum_{k=1}^{m_{pos}} R_k - \frac{m_{pos}(m_{pos} + 1)}{2} \right) \quad (12)$$

where m is the total number of images in the database. m_{pos} is the number of the ground truth in the database for a query and R_k is the rank of the k_{th} ground

Fig. 3. Comparison of the four similarity measures

Table 1. Comparison of the average rank value (S_1, S_2 : proposed measures)

Measure	S_1	S_2	S_3	S_4
avg_rank	15.71	15.83	17.13	15.99 ($\times 10^{-2}$)

truth image for a query. The smaller this average rank value is, the better the retrieval performance is.

The *Precision* and *Recall* curves are plotted in Figure 3. The horizontal axis is the number of retrieved images, and the top 50 retrieved images are evaluated. As shown in sub-figure(a), the proposed similarity measures, S_1 and S_2 , achieve better retrieval performance than S_3 and S_4 . The main difference between the two proposed measures and those in [1,2] lies in our new weighting scheme and the unnormalized inner product, and this result verifies their effectiveness. In addition, S_2 outperforms the other three at the first retrieved image, whereas S_1 becomes the best when the number of retrieved images increases. This shows that taking the number of occurrences of a visual word into account helps to identify a perfect matching database image which has many copies of queried objects.

Fig. 4. Some retrieval examples. The number of ground truth is listed in bracket.

However, this also makes the similarity measure more sensitive to noise because when a database image doesn't really contain the query object but contains multiple copies of another kind of object which shares one or two types of visual word with query object, it might score even more than those true matching image who has only one copy of query object through S_2 . In other words, S_2 focus on how many visual words query and database images have in common, when S_1 focus on how many types of visual word query and database images have in common. *Recall* is shown in sub-figure(b), from which similar conclusion can be drawn. In the sub-figure(c), the horizontal axis is the number of retrieved images, and the vertical axis shows the percentage of the queries for which at least one correct match is found. As it shows, through S_1 and S_2 , over 70% of the query images can find at least one correct match in top 50 retrieved images. However, through S_3 and S_4 , this number is only between 50% and 60%. This demonstrates again that the proposed S_1 and S_2 gain better performance than S_3 and S_4 do in our retrieval task.

The average ranks are listed in Table 1. It can be seen that the proposed S_1 produces the lowest value. It means that, averagely, the ground truth images are assigned higher ranks under this measure. This coincides with the *Precision* and *Recall* result. Some examples of the top 3 retrieved images are shown in Figure 4.

6 Conclusion

A new retrieval problem is proposed in this paper. The experimental result demonstrates the better performance of the proposed similarity measures. Meanwhile, it can be seen that less than half of the 1st retrieved images are correct answers and that only about 40% of true matches can be found after retrieving twenty images. Such a result indicates that this retrieval task is quite challenging and more work such as the verification of matches and the consideration of feature dependency needs to be explored to further boost the retrieval accuracy.

Acknowledgments. Many thanks to the Coles store located in Woden Plaza, ACT. We were allowed to collect the images in their store for our research. Their understanding and support make this work possible and are highly appreciated.

References

1. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, pp. 1470–1477. IEEE Computer Society Press, Los Alamitos (2003)
2. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2161–2168. IEEE Computer Society Press, Los Alamitos (2006)
3. Lowe, D.G.: Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision, pp. 1150–1157. IEEE Computer Society Press, Los Alamitos (1999)

4. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 63–86 (2004)
5. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. *International Journal of Computer Vision* 65(1/2), 43–72 (2005)
6. Lowe, D.G.: Local feature view clustering for 3d object recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 682–688. IEEE Computer Society Press, Los Alamitos (2001)
7. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 530–535 (1997)
8. Duda, R.O., Stork, D.G., Hart, P.E.: *Pattern Classification*, 2nd edn. vol. 115. John Wiley and Sons, Chichester (2001)