Image Retrieval with a Visual Thesaurus

Yanzhi Chen, Anthony Dick and Anton van den Hengel

School of Computer Science
The University of Adelaide
Adelaide, Australia

{yanzhi.chen, anthony.dick, anton.vandenhengel}@adelaide.edu.au

Abstract—Current state-of-art of image retrieval methods represent images as an unordered collection of local patches, each of which is classified as a “visual word” from a fixed vocabulary. This paper presents a simple but innovative way to uncover the spatial relationship between visual words so that we can discover words that represent the same latent topic and thereby improve the retrieval results. The method in this paper is borrowed from text retrieval, and is analogous to a text thesaurus in that it describes a broad set of equivalence relationship between words. We evaluate our method on the popular Oxford Building dataset. This makes it possible to compare our method with previous work on image retrieval, and the results show that our method is comparable to more complex state of the art methods.

Keywords—Image retrieval, image search, bag of words

I. INTRODUCTION

Image search and retrieval from large image collections has been a topic of much research in recent times, motivated by the availability of such collections on the web. The idea is to enable “true” image search based on the contents of each image rather than surrounding text or tags.

The current state of the art in searching for images of a specific object within a large database uses a number of techniques borrowed from textual information retrieval (IR). Among them, Sivic and Zisserman [8], [9] applied the bag-of-words method to the image search problem.

The approach begins with a feature detection process, which runs over the set of images to identify local patches of interest. Each feature is then represented as a descriptor vector, and images are represented as collections (“bags”) of descriptors. These vectors are then grouped into clusters called visual words. A visual vocabulary is formed as the collection of all the visual words, and every feature in every image is assigned one word. Following the previous steps, images can be represented as a weighted histogram (or vector) of visual word frequencies.

Following on from [8], a number of improvements have been proposed, including [3], [4], [6], [5]. Philbin et al. proposed [5] spatial verification to remove highly ranked false positives from the top ranked results by checking that the spatial layout of matching features was consistent. Chum et al.[3] proposed to improve retrieval performance by including more words in the query. This query expansion is based on finding more true positive matches, while relying on spatial verification to reject false positives.

Philbin et al.[6] implements query expansion by identifying individual features with multiple visual words in the vocabulary. The features in query images are no longer hard assigned to one cluster center while in the processing of quantization in forming a word vocabulary. Instead, each feature is assigned multiple nearest cluster centers. The soft assignment method uses a weighted query histogram bins instead of incrementing only the bin corresponding to the closest cluster center as was previously the case.

As a popular image representation method, bag-of-words has many successful implementations in image searching[8], [9], [3], [4], [6], [5] and categorization [7]. However, these works do not consider the spatial relationship between features when forming the vocabulary. We aim to introduce information as to how the features are spatially related to each other, and to utilize this spatial information to discover synonym relationships between words.

The idea of this paper is to perform query expansion using the spatial layout of features in images. Rather than try to discover similar feature vectors in feature space, we instead attempt to uncover the relationships between features by analysis of their co-occurrence in the images. The relationships we uncover through this method are similar to synonyms in text. The result is analogous to a text thesaurus in that it describes a broad set of equivalence relationships between words. We name the method described above as Visual Thesaurus. By uncovering these relationships, we can add spatial information to early steps of retrieval and improve the performance of the search process.

The motivation of our visual thesaurus approach is that features repeatedly appear and disappear as a set of small transformations are applied to an image. The images in Figures 1 show some of the MSER[11] features for a set of images which differ from each other by a slight projective transformation. These images illustrate the problem, with a number of features reliably detected throughout the image set, but a significant set of features which are detected in only a subset. The visual thesaurus uses the stable features to learn the relationships between the unstable ones. These relationships are useful when the query patch has a large number of features, but critical when only a small number
of features are present. The visual thesaurus differs from soft assignment in that it is based on spatial co-occurrence of features in images rather than neighbourhood relationships in feature space. Thus although both methods are a form of query expansion they operate in different ways. The visual thesaurus also has the advantage that it can work directly on the visual words rather than requiring that the features be clustered again in order to record multiple nearby cluster centers.

Figure 1: The output of the MSER detector when applied to 2 almost identical images. The first image is the original and the second image generated by applying slight projective warps. Note that although many of the detected features are very stable, a significant number of them also appear, disappear, and mutate between images.

The rest of the paper is structured as follows. In Section II, we describe the offline process of forming a visual thesaurus for an image corpus. In Section III, we describe how the thesaurus is used to perform image retrieval. The result of using a thesaurus is compared to a baseline method and to soft assignment in Section IV, including overall results and specific success and failure cases. We summarise the paper in Section V.

II. FORMING A VISUAL THESAURUS

The basic idea of a visual thesaurus is to measure the co-occurrence of individual image features. We use a set of histograms to record the co-occurrence information, so that for each word in the vocabulary, a histogram is created to record its co-occurrence with every other word in the vocabulary. The details are as follows: Nearest neighbors: In order to form the thesaurus, a sorted nearest neighbor list is computed for each feature point. The nearest neighbors are calculated in image space, and the algorithm loops over each feature points in every image of the corpus. The distances between the feature points are simply measured by the Euclidean distance. The number of entries in the list can be fixed, or based on a distance threshold. Forming the Thesaurus: The thesaurus is made up of a set of histograms counting the number of times each pair of features occurs in each other’s nearest neighbour list. Note that the size of the set of histograms corresponds to the number of the words in the vocabulary. In detail, we first identify the word which each feature belongs to, and then iterate over the nearest neighbors of the feature point. If the closest neighbor exists, the corresponding word in the vocabulary is retrieved and the entry of the histogram for this feature point is incremented. The process is repeated for every feature point in all images. After accumulation all histograms are normalised, and the majority of entries are zero. In some cases, however, there is a correlation between the occurrence of features. This leads to a peak in one or more of the histogram elements beyond that predicted by the density of the features in the database. We define the set of histograms based on nearest neighbours as a \( \{ F \} \) thesaurus. This reflects that we generate a histogram for every word in the vocabulary and increment the count in that histogram corresponding to the next (single) closest feature point. With more feature points from the nearest neighbour list included, the thesaurus can enlarge the region of neighbors. Under such cases, we define an \( \{ F \} \) thesaurus for recoding the \( F \) closest features. Suppose we know the corpus \( \mathcal{D} \) with M terms \( \{ d_1, d_2, \ldots, d_M \} \) containing visual words \( \mathcal{P} \) in the vocabulary. The visual words in query region can be expressed as \( \mathcal{Q} = \{ q_1, q_2, \ldots, q_n \} \), where \( n \) is the number of query words and \( \forall q_n \in \mathcal{P} \). Define \( I \) as an index function, we link the features \( f \) and the word \( q \) with the condition that \( q = I(f) \) so that the word \( q \) is the feature \( f \) belongs to. The histograms for the words in the vocabulary is denoted as \( H_I \). The pseudo code for the formation of an \( \{ F \} \) thesaurus is given in Algorithm 1:

**Algorithm 1** Forming an \( \{ F \} \) thesaurus.

```plaintext
for all images \( \{ a \} \) in \( \mathcal{D} \) do
    for all features \( \{ f \} \) in \( a \) do
        retrieve the histogram \( H_I(f) \)
        for all \( F \) closest features \( \{ g \} \) do
            Increment the element \( I(g) \) of \( H_I(f) \)
        end for
    end for
end for

for all words \( \{ q \} \) in the vocabulary \( \mathcal{P} \) do
    Normalise s.t. \( \sum_q H_q = 1 \) \( \forall q \in \mathcal{D} \)
end for
```

Our method in this paper relies on the highest entries of each histogram formed as thesaurus, and will implement the experiments in Section 4.

III. RETRIEVAL WITH VISUAL THESAURUS

The purpose of visual thesaurus is to add extra words \( C = \{ c_1, c_2, \ldots, c_m \} \) from \( P \) to the query. The retrieval process utilizes the extra words \( C \) to find the latent objects and improve the retrieval results. After the extra words have
been found, we just add them to the the original query words set:

\[ Q^* = [Q|C] = [q_1, q_2, \ldots, q_n, c_1, c_2, \ldots, c_m]; \]  

(1)

There are two ways to judge whether an \( I(g) \) in Algorithm 1 should be included in the new query \( Q^* \): by counting the co-occurrence in the histograms or by thresholding the normalized histograms in the thesaurus. Algorithm 2 shows that if \( I(g) \) is greater than the threshold \( t \) then this word will be added to the query \( Q^* \). In the same way, if the counted co-occurrence used, the query \( Q^* \) is expanded by words occurring most frequently in \( H_I(q) \). Although it is simpler to implement the first method, the second approach always performs better that the first one in our experiments, since the process of normalization can avoid peaks in histograms which are unrelated. We ignore the first kind of threshold for the rest of this paper. The algorithm 2 shows the process of adding extra words to the query \( Q^* \). More details will be discussed in Section 5 and Section 6, including the setting of parameters and the way of choosing the threshold.

Algorithm 2 Query with an \( \{ F \} \) thesaurus.

```plaintext
for all words \( \{ q \} \) in the query region \( Q \) do
  retrieve the histogram \( H_I(q) \)
  for all words \( c \) in \( H_I(q) \) do
    if \( H_I(q)(c) > t \) then
      Add \( c \) to \( C \).
    end if
  end for
end for
Query with \( Q^* = [Q|C] \).
```

IV. EXPERIMENTS

We test and illustrate our method for image retrieval using the popular Oxford building dataset [2]. The vector-space search engine we use in this paper is the same as in [8], [9]. The search process compares the query vector and the target vectors by the distribution of visual words found in the images. This uses the standard tf-idf term weighting [10]. We use the Average Precision (AP) [9] for evaluation and comparison with base line methods. The mAP value refers to the area under the Precision-Recall curve as in previous work [6], [3], [5]. Following [3], each of the 5 queries of a landmark has an Average Precision score, and we average them to compute a Mean Average Precision (mAP) for each landmark. The total mAP value is computed from the 11 landmarks’ mAP scores. Images are labeled manually as Good, OK, Junk, and Bad. The performance is evaluated using the “Good” and “OK” as true positive, “Bad” as negative. The “Junk” images are treated as unaffected ground truth, which will not be counted as positives or negatives. The evaluation will be used in the following of the experiments in this paper.

**Base line:** The base line method is defined as retrieval with the standard tf-idf term computed from the corpus, to weight the similarity between the query vector and each document vector. The approach only retrieves based on words found in the query region: \( Q = [q_1, q_2, \ldots, q_n] \). No spatial information considered in the base line method.

**Visual Thesaurus:** Our experiments start with a \( \{ 1 \} \) thesaurus, i.e. using only the single closest features, and then add more neighbouring features if they exist. Table I shows details of 6 kinds of thesaurus query expansion. In this step, the choice of threshold \( t \) for inclusion in the thesaurus is important as mentioned in Section III. The value of \( t \) we use in this paper is 0.1 for the \( \{ 1 \} \) threshold, and this decreases by 30% each time the number of neighbour features is increased. For instance, the threshold value in \( \{ 1 \} \) thesaurus is 0.1, which gives the best output. Then the value decrease to be 0.07 in \( \{ 2 \} \) thesaurus, and 0.049 in \( \{ 3 \} \) thesaurus, so far so on. The approach retrieves with \( Q^* = [Q|C] \).

Spatial Verification: The base line method gives a ranked list of images in the corpus. However, there might be high ranked false positives due to lack of consideration of spatial information. Spatial verification helps to filter out the false positives by estimating a transformation between the query and target images. In this paper, we implement spatial verification by estimating a 6 dof (degree of freedom) transformation from the inliers of a simpler 3 dof one with at least 3 inliers included. Note that the process of verifying a visual thesaurus involves \( Q^* = [Q|C] \), with more feature information compared with base line method. We match the features assigned to same words from \( Q \) first. If the corresponding \( q \in Q \) can not be found in the target images, we try to search the synonyms \( C \) instead. The results are re-ranked based on the sum the scores of tf-idf weighting and the number of inliers if the transformation can be estimated. Note that our version of spatial verification is slightly different than [6], [5], but the results are very similar. We will illustrate the results of spatial verification in following part.

A. Retrieval performance comparison

We compare our results on the Oxford building dataset [2] with the base line methods and our 6 kinds of thesaurus expansion. The dataset includes 5,062 images from the photo-sharing website Flickr [1], with a set of 16,334,970 detected features and their cluster centers supplied on their website [2]. We demonstrate detailed results on all souls l and hertford l.

Experiments in table I show the 11 Oxford building landmarks’ mAP scores. The mean average precision score is better if more neighbour features are included in the thesaurus. It reaches the highest score when the number of the nearest neighbors is defined as \( r <= 6 \), forming a \( \{ 6 \} \)
Figure 3: Precision recall curves for all souls 1 and hertford 1 with base line method and 6 thesaurus extension.

The precision recall curves of the two particular case all souls 1 and hertford 1 are illustrated in figure 3, which show the improvement of the 6 kind of thesaurus listed in table I.

Figure 2 shows an example of feature correspondence between the query image and candidate target images, with and without the synonyms introduced by the visual thesaurus. Correspondence is based on the features assigned to the same word. In these correspondence, most of the extra words occur in the right place—the extra words are located near where the query words \( Q = [q_1, q_2, \ldots, q_n] \) detected—and can be viewed as a supplement to \( Q \).

Table II compares the results from [6] with our experiments. The expression \( r = 0 \) means no other information has been included, so both Soft Assignment and Visual Thesaurus reduce to the baseline method. Under the same condition, the mAP value of \( \{5\} \) thesaurus (with 5 closest neighbors) is greater than soft assignment with 5 nearest cluster centers, and can improve more with \( \{6\} \) thesaurus.

The effect of spatial verification for the base line method and \( \{5\} \) thesaurus are listed in table III. The re-ranking is implemented for the top 400 results, with at least 3 inliers required to estimate the transformation. We illustrate the verified results of 3 kind of methods: base line, \( \{1\} \) thesaurus and \( \{5\} \) thesaurus in figure 4. The thesaurus helps to find more correspondence in images and thus locate the objects more accurately.

We stop our experiment at \( \{6\} \) thesaurus in Table I. The spatial co-occurrence of features in the extra words \( C \) is less likely to be meaningful as the region of closest feature becomes larger. The visual thesaurus relies on these extra words \( C \) to uncover the relationship between features. In this paper we decrease the threshold \( t \) while increasing \( r \), so that the thesaurus becomes more selective as the number of neighbours included grows. Table IV analyzes the extra words introduced by different versions of the visual thesaurus. To trade off the precision and the computation time, we find it is better to decrease the threshold \( t \) linearly as \( 0.7(t_1) \).

### B. Example results

Figure 5 shows examples of retrieved facades from the 5,062 images, comparing the baseline method with the \( \{5\} \) thesaurus. All the shown results are more highly ranked than the first false positive. The visual thesaurus helps to find more high ranked positives by querying with \( C \), and therefore locates the object more accurately.

However, our method does not always perform well. Some high ranked false positive still can not be moved from the top of the rank list. Some of the errors are shown in Figure 6. The left hand case is hard to overcome due to the fact that ambiguities occur in the visual words in the images. Note that there are a lot of extra words included by the thesaurus, which means it create many extra pairs of matched feature points. This causes difficulty for spatial verification to tell which transformation is "true". The

<table>
<thead>
<tr>
<th>number of nearest cluster centers</th>
<th>( r = 0 )</th>
<th>( r = 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Assignment</td>
<td>0.614</td>
<td>0.673</td>
</tr>
<tr>
<td>Visual Thesaurus</td>
<td>0.614</td>
<td>0.685</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>number of nearest neighbour features</th>
<th>( r = 0 )</th>
<th>( r = 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft Assignment</td>
<td>0.614</td>
<td>0.673</td>
</tr>
<tr>
<td>Visual Thesaurus</td>
<td>0.614</td>
<td>0.685</td>
</tr>
</tbody>
</table>

Table II: Comparison between soft-assigned vocabularies and visual thesaurus.

<table>
<thead>
<tr>
<th>BL</th>
<th>BL + SP</th>
<th>( TH{5} )</th>
<th>( TH{5} + SP )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.615</td>
<td>0.650</td>
<td>0.685</td>
<td>0.716</td>
</tr>
</tbody>
</table>

Table III: Spatial verification for the base line and \( \{5\} \) thesaurus methods.

1 Due to the RANSAC processing and the difference between our version of spatial verification and [6], we compare the improvement based on our experiments.
Figure 2: Examples of queries *all souls* 1 and *hertford* 1. (a): feature correspondence with $Q$ without the synonyms introduced by the visual thesaurus. (b): extra feature correspondence with synonyms introduced by {1} thesaurus, only show features with $C$. The thesaurus query with features both in (a) and (b).

Figure 4: Examples of verified searching results of *all souls* 1 and *hertford* 1. (a): spatial verification of base line method; (b): spatial verification of {1} thesaurus; (c): spatial verification of {5} thesaurus.
Table I: Evaluation for the 11 Oxford building landmarks. The values are shown in percentage, except the total mAP values computed from all the 11 landmarks. The top line characters represent the methods we use in the experiment: "BL" - baseline, "TH\{1\}" - \{1\} thesaurus, "TH\{2\}" - \{2\} thesaurus, "TH\{3\}" - \{3\} thesaurus, "TH\{4\}" - \{4\} thesaurus, "TH\{5\}" - \{5\} thesaurus, "TH\{6\}" - \{6\} thesaurus.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Methods</th>
<th>Good + OK</th>
<th>Junk</th>
<th>BL</th>
<th>TH{1}</th>
<th>TH{2}</th>
<th>TH{3}</th>
<th>TH{4}</th>
<th>TH{5}</th>
<th>TH{6}</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Souls</td>
<td>78</td>
<td>111</td>
<td></td>
<td>53.33</td>
<td>60.30</td>
<td>65.31</td>
<td>66.51</td>
<td>67.30</td>
<td>69.36</td>
<td>71.08</td>
</tr>
<tr>
<td>Ashmolean</td>
<td>25</td>
<td>31</td>
<td></td>
<td>59.57</td>
<td>61.93</td>
<td>62.67</td>
<td>64.15</td>
<td>63.77</td>
<td>64.09</td>
<td>64.49</td>
</tr>
<tr>
<td>Balliol</td>
<td>12</td>
<td>18</td>
<td></td>
<td>59.40</td>
<td>62.57</td>
<td>62.97</td>
<td>60.11</td>
<td>63.60</td>
<td>64.37</td>
<td>64.43</td>
</tr>
<tr>
<td>Bodleian</td>
<td>24</td>
<td>30</td>
<td></td>
<td>56.77</td>
<td>57.78</td>
<td>59.40</td>
<td>60.13</td>
<td>62.16</td>
<td>61.38</td>
<td>62.05</td>
</tr>
<tr>
<td>Christ Church</td>
<td>78</td>
<td>133</td>
<td></td>
<td>62.28</td>
<td>69.13</td>
<td>72.23</td>
<td>71.84</td>
<td>72.29</td>
<td>72.49</td>
<td>72.43</td>
</tr>
<tr>
<td>Cornmarket</td>
<td>9</td>
<td>13</td>
<td></td>
<td>56.31</td>
<td>58.80</td>
<td>57.11</td>
<td>57.28</td>
<td>58.40</td>
<td>59.05</td>
<td>57.87</td>
</tr>
<tr>
<td>Herford</td>
<td>24</td>
<td>31</td>
<td></td>
<td>76.43</td>
<td>80.5</td>
<td>81.61</td>
<td>82.43</td>
<td>84.26</td>
<td>84.65</td>
<td>78.82</td>
</tr>
<tr>
<td>Keble</td>
<td>7</td>
<td>11</td>
<td></td>
<td>69.95</td>
<td>66.32</td>
<td>70.25</td>
<td>69.62</td>
<td>73.04</td>
<td>80.17</td>
<td>80.50</td>
</tr>
<tr>
<td>Magdalen</td>
<td>54</td>
<td>103</td>
<td></td>
<td>15.47</td>
<td>16.13</td>
<td>17.38</td>
<td>17.26</td>
<td>15.65</td>
<td>15.20</td>
<td>13.43</td>
</tr>
<tr>
<td>Pitt River</td>
<td>7</td>
<td>9</td>
<td></td>
<td>59.48</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Radcliffe Cambridge</td>
<td>221</td>
<td>348</td>
<td></td>
<td>67.43</td>
<td>74.72</td>
<td>77.51</td>
<td>76.41</td>
<td>81.40</td>
<td>82.59</td>
<td>83.27</td>
</tr>
<tr>
<td>Total</td>
<td>539</td>
<td>838</td>
<td></td>
<td>0.614</td>
<td>0.645</td>
<td>0.66</td>
<td>0.662</td>
<td>0.674</td>
<td>0.685</td>
<td>0.689</td>
</tr>
</tbody>
</table>

Table IV: The first row shows the number of extra words introduced by different thesauri. Experiments begin with TH\{1\} and a threshold \(t = 0.1\) as described above. Two types of threshold \(t_j\) and \(t_{j+}\) represent step decrease by 30% and 20% in each time the thesaurus is changed from TH\{2\} to TH\{6\}. The second row shows the corresponding mean average precision value.

right images is a "successful" verification, although it also contains lots of wrong matching. It can be distinguished to be a false retrieved results since no transformation could be found in verification. The picture originally appears in the 3rd place in the base line method, then the \{5\} thesaurus helps to move it to a lower place at 8th and finally drops to 39th in the rank list after spatial verification used in \{5\} thesaurus.

![Examples of the first high ranked false positives of \{1\} thesaurus in: Ashmolean and All Souls. The examples are high ranked false positives from the top 100 of the retrieval results.](image)

V. CONCLUSION AND FUTURE WORK

The visual thesaurus captures information latent in an image database and uses it to improve the performance of specific object image search. The process can be divided into off-line and on-line steps and is relatively simple to implement and use. We demonstrate our idea on the Oxford building datasets and reach the conclusion that the \{5\} thesaurus is the best trade-off between efficiency and accuracy in this case. Testing on the Oxford buildings database shows that our method can improve overall retrieval results and outperform related methods under the same conditions. The \{5\} thesaurus method improved the average precision of the bag-of-words approach from 0.615 to 0.685, compared with the soft assignment process proposed in [6] achieves 0.673 average precision in the same circumstances. Additionally, the thesaurus method can perform better than base line retrieval using spatial verification, without re-ranking processing.

It should be noted that the extra words included by the visual thesaurus and their effect on retrieval is the main focus of this paper. It may also be worthwhile to classify these extra words into different topics, but this is not completed in this paper. We leave this as feature work.
ACKNOWLEDGMENT

The authors would like to thank discussion with Zygmunt Szpak, Hoi Sim Wang and Liang Li, and help from Rhys Hill in the experiments.

REFERENCES


