Pruning near-duplicate images for mobile landmark identification: a graph theoretical approach

Taner Danisman, Jean Martinet, Ioan Marius Bilasco
CRISTAL, Lille 1 University, Villeneuve d’Ascq, France
Email: \{taner danisman, jean.martinet, marius.bilasco\}@univ-lille1.fr

Abstract—Automatic landmark identification is one of the hot research topics in computer vision domain. Efficient and robust identification of landmark points is a challenging task, especially in a mobile context. This paper addresses the pruning of near-duplicate images for creating representative training image sets to minimize overall query processing complexity and time. We prune different perspectives of real world landmarks to find the smallest set of the most representative images. Inspired from graph theory, we represent each class in a separate graph using geometric verification of well-known RANSAC algorithm. Our iterative method uses maximum coverage information in each iteration to find the minimum representative set to reduce and prioritize the images of the initial dataset. Experiments on Paris dataset show that the proposed method provides robust and accurate results using smaller subsets.

I. INTRODUCTION

An increasing number of tourist applications use landmark identification, in addition to GPS coordinate, to provide users with contextual informations. Landmark identification is fundamentally a generalization of the object identification problem. With the advancements in mobile technologies, there is a growing demand for creating effective landmark identification applications using state-of-the-art methodologies.

Recent literature in both object and landmark identification includes local feature based methods (e.g SIFT, SURF, BRIEF) used with an appropriate classifier [Lowe, 2004] [Sivic and Zisserman, 2003]. RANSAC algorithm [Fischler and Bolles, 1981] is also used to verify the geometric constraints of the object. The algorithm accepts or rejects a given scene image with respect to an object image by considering the geometric layout of paired keypoints. This method was originally designed to identify planar objects, therefore its performance is affected when a 3D object of interest is presented from different view points. More challenge comes from external factors such as daytime/night-time illumination change, dynamic object occlusions (e.g. vehicles, trees, people), rotation, and scaling. In this scope, Bag of Words (BoW) methodology is well adapted to visual object and scene recognition [Sivic and Zisserman, 2003]. However, it needs a high dimensional visual dictionary to identify a given query image.

All of the aforementioned methods use large number of training images for processing, clustering and identification. Therefore, there is a need to summarize these large volumes of information. Variations originating from the same object are common in image collections. The most obvious example of this type of redundancy is duplicate or similar images. When dealing with touristic landmarks (cityscapes, monments, famous buildings, etc.), and due to the vast amount of available images, there is a need to prune this redundant information. Near-duplicate images are groups of similar images that vary slightly because of different camera view positions, seasonal changes, scale and orientation. Previously, elimination of near-duplicate/co-derivative images studied in [Foo et al., 2006] using relationship graphs and locality-sensitive hashing [Gionis et al., 1999]. They applied refine-and-filter scheme by first pruning using a hash table, where PCA-SIFT features sharing identical hash-keys are stored in the same entry. They updated a relationship graph with respect to the co-occurrence of the patches and units. However, they did not considered the geometrical verification of the object of interest.

The main contribution of this study is the improvement in query response time by pruning the redundant information using a graph theoretical approach. Instead of feature level-matching, our proposed method works at the image level. As a result, it provides better compression and de-noising of initial set. The remainder of this paper is organized as follows. Section 2 presents the methodology covering landmark matching and pruning near-duplicate images with detailed examples. Section 3 describes the experimental setup, evaluation metrics and results. Finally, section 4 summarizes the approach, brings a discussion and concludes the paper.

II. METHODOLOGY

In object recognition, high-dimensional spaces for object representation are often sparse, which prevents the identification process from being efficient. This problem is referred to as the curse of dimensionality. Elimination of near-duplicate images in object recognition is a required step targeting redundancy reduction, while preserving characteristic views of the objects. As an interesting side effect of our approach, potential outliers are automatically eliminated. In other words, it provides data compression and de-noising. Since it is well known that the most important intrinsic variables are rotation, translation and scale, we use the methods targeting these aspects of object recognition.

A. Landmark Matching

First of all, all images inside a given class are individually matched against the rest of the images from the same class. Considering an n-class object identification problem, a complete object match is performed for each class \(C_p\) from our dataset where \(p = \{1, ..., n\}\) is the class number among \(n\). For the matching process, a common object matching chain is used, starting with the keypoint detection, feature extraction,
relevant feature selection and finally geometric verification steps. Within a given class, 64-bin original SURF features are used to match one image \( I_i \) against the rest of the images \( I_j \) in the class, where \( i \neq j \) and \( C_p = \{ I_1, I_2, ..., I_m \} \) is the set of \( m \) images in the class. Geometrical verification is performed using RANSAC-based homography to find the translation and rotation matrix of the detected object. In order to eliminate matching errors, convex hull and line slope tests are applied on the homography results. Note that this step also helps focusing on points lying inside the object of interest, since they are more likely to match one another. It is possible to accept or to reject a given set of points from \( C_p \). Therefore, a binary coverage vector \( \vec{v}_i \) of order \( m \) can be generated for each query image \( I_i \), where all correct matches are represented by 1 and no match results are represented by 0. The set of \( \vec{v}_i \) generated from all \( I_i \) generates the \( V_p \) matrix of size \( m \times m \) where each \( v_{ij} \) denotes the match between images \( I_i \) and \( I_j \). Self-matching (diagonal elements in \( V_p \)) is obviously not considered and therefore is disabled by setting all \( v_{ii} = 0 \), like in the set of vectors \( V_p = \{ \vec{v}_1, \vec{v}_2, ..., \vec{v}_7 \} \) given Eq.(1) as an example. Let \( f(\vec{v}_i) \) be the sum of all elements in a given row vector \( \vec{v}_i \), then value of \( f(\vec{v}_i) \) presents the coverage of image \( I_i \) in class \( C_p \) (see Eq.(2)). A consensus set \( R_p \) is iteratively built as described in the next section.

\[
\vec{v}_1 = [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0] \\
\vec{v}_2 = [1 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0] \\
\vec{v}_3 = [0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0] \\
\vec{v}_4 = [0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0] \\
\vec{v}_5 = [0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0] \\
\vec{v}_6 = [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1] \\
\vec{v}_7 = [1 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0] \\
\]

\[
f(\vec{v}_i) = \sum_{k=1, \vec{v}_k \not\in R_p}^{m} v_{ik} \tag{2}
\]

### B. Pruning near-duplicate images

In order to select only representative images, we wish to eliminate near-duplicate images from the initial set. Figure 1 shows near-duplicate images of Arc de Triomph class from Paris dataset [Philbin et al., 2008] and corresponding relative camera positions in cartesian coordinate system using VisualSfM [Wu, 2013]. Using all images for the identification process provides multiple detections of the objects, as shown Figure 2. However, this first approach is both computationally complex and time consuming.

Considering a set of images belonging to a specific class \( C_p \), our algorithm first searches for the \( \vec{v}_i \) that maximizes the \( f(\vec{v}_i) \) as an initial consensus set, as described in Eq.(3).

\[
R_p = \{ \vec{v}_i | i = \arg\max_k f(\vec{v}_k) \} \tag{3}
\]

Figure 3 shows the graph representation of \( V_p \) where marked dark vertex points are the members of \( R_p \). Formally, this problem is equivalent to the minimum edge cover problem in graph theory, and from a clustering perspective, the main idea is similar to that of Highly Connected Subgraphs (HCS) clustering [Hartuv and Shamir, 2000]. However, we ignore the edges \( v_{ij} \) where both the vertex \( i \) and vertex \( j \) are matched by any vertex from \( R_p \). Since minimal edge covering is an optimization problem that can be solved in a polynomial time, we used greedy maximum matching so that all vertices are covered. A vertex that is already matched by a selected vertex from the \( R_p \) is simply ignored. For the particular example in Eq.(1), \( \vec{v}_2 \) provides the maximum coverage value \( f(\vec{v}_2) = 4 \). The next iteration searches for a \( \vec{v}_j \) such that a new vector \( \vec{v}' = \vec{v}_2 \lor \vec{v}_j \) maximizes \( f(\vec{v}') \) in the set \( V_p \). Therefore, the next iteration selects \( \vec{v}_5 \) since \( \vec{v}' = \vec{v}_2 \lor \vec{v}_5 = [1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1] \) and \( f(\vec{v}') = 6 \). These iterations continue to select the next most representative image from \( V_p \) until there is no more change in the maximum value of \( f(\vec{v}') \). Note that \( f(\vec{v}_7) > f(\vec{v}_5) \) when \( R_p = \{ \vec{v}_2 \} \). \( \vec{v}_5 \in R_p \) makes \( f(\vec{v}_7) = 2 \) instead of 3, because of the constraint \( \vec{v}_k \not\in R_p \) from Eq.(2). As a result, the final representative set \( R_p \) for the class \( C_p \) includes only \( \vec{v}_2 \) and \( \vec{v}_5 \): \( R_p = \{ \vec{v}_2, \vec{v}_5 \} \). Therefore, the final value of \( \vec{v}' \) which also represents the union of all \( \vec{v}_i \in R_p \) becomes \( \vec{v}' = [1 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1] \). The algorithm stops here since there is no more possibility to increase this value using existing vertices. The final step updates the arcs in \( \vec{v}'_{12} \) where \( \vec{v}' \in R_p \), which sets \( \vec{v}'_{12} = 1 \), \( \vec{v}'_{15} = 1 \) thus \( \vec{v}' = [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1] \) provides a complete coverage. For this particular example in Eq.(1), the
The set of all \( \{ R_p | p = 1, ..., n \} \) is used in our approach as a new training set. Only representative images are kept, making all subsequent processes lighter, while guaranteeing a high image coverage.

III. EXPERIMENTAL SETUP AND EVALUATION

We performed experiments on Paris dataset [Philbin et al., 2008] using six classes. Images are manually cropped to present only the object of interest as shown in Figure 5. For evaluating our method we used the coverage metric which is similar to the recall in document retrieval literature [Foo et al., 2007]. In addition, we performed timing experiments including feature extraction, landmark matching and query response time.

For the timing experiments, we compute the average query time and speed-up in the original and pruned set. There are several factors that affects the query response time. First, minimum hessian distance parameter that is used by the keypoint detection function. It affects the total number of keypoints that is transferred to the next step. Second, the amount of important keypoints that are transferred to RANSAC. This is determined by a confidence value that correspond to the similarity ratio between the two closest features. Similar to [Philbin et al., 2007], we used 4 randomized k-d trees to speed-up the nearest neighbour computations between the points and cluster centers. The two closest features satisfying a given confidence value are selected as a good match. We empirically choose 0.6 for the confidence value. All experiments are performed on an Intel Xenon 3.0 GHz processor without any parallel execution setting.

Figure 6 presents result of the pruning for Arc de Triomphe class, where \( RF_j \) presents the set of representative images in ranked order. The representative set consists of five images covering all the dataset. Table I shows the coverage and timing information for individual processing steps. In order to see the discriminative power of the representative set \( R_p \), we performed tests on the negative classes \( C_i, i \neq p \). Table II shows the average query time and relative speed-up and precision per class considering the negative tests. Table III presents the confusion matrix, that mainly consists of only diagonal elements without false positives. Since the algorithm cannot decide on some cases, an Unknown class is added to the confusion matrix. For the evaluations, the Unknown class is used as an equivalent to false positives.

According to Table I, Tour Eiffel class has the lowest coverage value of 69.57%. The main reason comes from the architectural background of the Eiffel tower. The holes inside the iron structure provide many different possibilities that the feature extraction step generates drastically different features. Because of this unique property, we need more representative images to cover individual images. A similar issue stands for
TABLE II.  SUMMARY OF THE EXPERIMENTS. AQT=AVERAGE QUERY TIME (SECONDS), SU=SPEED-UP, NTS=NEGATIVE TEST SIZE, P=PRECISION.

<table>
<thead>
<tr>
<th>Class</th>
<th>AQT</th>
<th>SU</th>
<th>NTS</th>
<th>P (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arc de Triomphe</td>
<td>0.88</td>
<td>6.62</td>
<td>635</td>
<td>98.49</td>
</tr>
<tr>
<td>Tour Eiffel</td>
<td>0.59</td>
<td>3.27</td>
<td>721</td>
<td>96.9</td>
</tr>
<tr>
<td>Les Invalides</td>
<td>1.45</td>
<td>3.76</td>
<td>613</td>
<td>100</td>
</tr>
<tr>
<td>Moulin Rouge</td>
<td>2.55</td>
<td>1.50</td>
<td>541</td>
<td>100</td>
</tr>
<tr>
<td>Sacré cœur</td>
<td>1.65</td>
<td>2.88</td>
<td>654</td>
<td>100</td>
</tr>
<tr>
<td>Notre Dame</td>
<td>1.09</td>
<td>5.18</td>
<td>671</td>
<td>100</td>
</tr>
</tbody>
</table>

TABLE III.  CONFUSION MATRIX PRESENTING NUMBER OF CORRECTLY IDENTIFIED IMAGES. ADT=ARC DE TRIOMPHE, TE=TOUR EIFFEL, LI=LES INVALIDES, MR=MOULIN ROUGE, SC=SACRE CŒUR, ND=NOTRE DAME, U=UNKNOWN.

<table>
<thead>
<tr>
<th>Class</th>
<th>ADT</th>
<th>TE</th>
<th>LI</th>
<th>MR</th>
<th>SC</th>
<th>ND</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADT</td>
<td>131</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>TE</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>LI</td>
<td>0</td>
<td>0</td>
<td>153</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>208</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>SC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>111</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>ND</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>92</td>
<td>4</td>
</tr>
</tbody>
</table>

Moulin Rouge images, where the moving sails yield different views of the object. Therefore, it has the second lowest coverage value of 92.04%. The rest of the coverage results showed that proposed method successfully pruned large amount of near-duplicate images from the initial set. Negative set tests also shows that RANSAC geometric verification and pruning step provide a high precision. Considering the negative class tests, we obtained very high precision due to the geometric verification.

IV. CONCLUSIONS

In this study, we present a graph theoretical approach for landmark identification by eliminating the near-duplicate images. The proposed method uses the well-know RANSAC algorithm for geometric verification of the landmarks. Inspired from the edge covering problem in graph theory, we show that proposed method provides a high coverage with a minimum set of representative images with high precision.

The presented study has considered six classes. We expect our approach to increase the scaling capability of nearest neighbour standard approaches, since the number of sample images is limited. However, we can expect the precision to decrease when the number of classes is increased. This issue will be studied in future works. In a more general direction, we plan to export the proposed method for city-level mobile landmark identification problem.

REFERENCES


