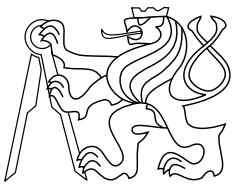
CENTER FOR
MACHINE PERCEPTION



CZECH TECHNICAL
UNIVERSITY IN PRAGUE

RESEARCH REPORT

ISSN 1213-2365

Web Scale Image Clustering

Large Scale Discovery of Spatially Related Images

Ondřej Chum and Jiří Matas

CTU–CMP–2008–15

May 23, 2008

ICT-215078 DIPLECS

Research Reports of CMP, Czech Technical University in Prague, No. 15, 2008

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Web Scale Image Clustering

Large Scale Discovery of Spatially Related Images

Ondřej Chum and Jiří Matas

CMP, Faculty of Electrical Engineering, Czech Technical University in Prague

Abstract. We propose a randomized data mining method that finds clusters of spatially overlapping images. The core of the method relies on the min-Hash algorithm for fast detection of so-called cluster seeds. The seeds are then used as visual queries to obtain clusters which are formed as transitive closures of sets of partially overlapping images that include the seed. We show that the probability of finding a seed for an image cluster rapidly increases with the size of the cluster.

The properties and performance of the algorithm are demonstrated on datasets with 10^4 and 10^5 images. The speed of the method depends on the size of the database and is close to linear for databases sizes up to approximately $2^{34} \approx 10^{10}$ images. The proposed algorithm provides, as a side effect, a state-of-the-art near duplicate image detection.

1 Introduction

Collections of images of ever growing sizes are becoming common both due to commercial efforts [1] and as a result of photo and video sharing of individual people [2, 3]. Structuring and browsing large images databases is a challenging problem. Developments like Photo Tourism [4] show that access to images based on the 3D location of the acquisition location or on the spatial overlap of the scenes they depict is intuitive and has high user acceptability. Commonly, the sets of relevant spatially related images are obtained using user annotations; we propose a method discovering spatial overlaps from image content alone using image retrieval techniques.

Recent image and object retrieval systems[1]support visual search even in large databases [5–9]. Starting from a visual example including an instance of the object of interest, such systems are able to retrieve relevant images with both high precision and recall. A direct application of this methodology to the data mining task would be to take in turn each image in the database and query the database with it. The method is quadratic in the size of the database[2] and hence not feasible for large databases.

In this paper, we propose a randomized data mining method to find clusters of images with spatial overlap; the probability of discovering a cluster is a function of its size and approaches one fast. Instead of trying to match each image in turn, a randomized procedure, the min-Hash, detects so-called cluster seeds. The seeds are then used as visual queries to obtain clusters which are formed as transitive closures of sets of partially overlapping images that include the seed. We show that the probability of finding a seed for an image cluster rapidly increases with the size of the cluster. For practical database sizes the running time of the seed generation process is close to linear in the the size of the database. The cluster completion process requires time proportional to the number of images in all clusters, which seems unavoidable.

The proposed unsupervised clustering method for large (web scale) image databases thus has the following desirable properties: it is (i) scalable – expensive operations, such as querying the whole database, are not applied to every single image, and are only applied to a number of images proportional to the number of images in the clusters (ii) incremental – adding new images into the database is possible without recomputing the whole clustering, (iii) the probability of discovering a cluster is independent of the database size.

How can the clustering for 3D modelling? So far, the fusion of the two tools – image retrieval and 3D registration systems – can be used. The user can choose an image of a place he or she is interested in and use a system that would virtually take him/her to that place. Using an image of that place as a visual query, the retrieval system looks up the photos of other people in the world. If there are enough photographs of this place available, a retrieval system collects the relevant images that are consequently used for the 3D reconstruction.

[1] By "object retrieval" we mean retrieval of a particular object (e.g. "my car"), not a class of objects (e.g. "a car"). We use the terms "categorization" or "object class recognition" for the latter problem.

[2] For retrieval systems based on inverted files, each query has to touch all images that have at least one visual word in common with the query image. The number of such images is proportional to the size of the database.

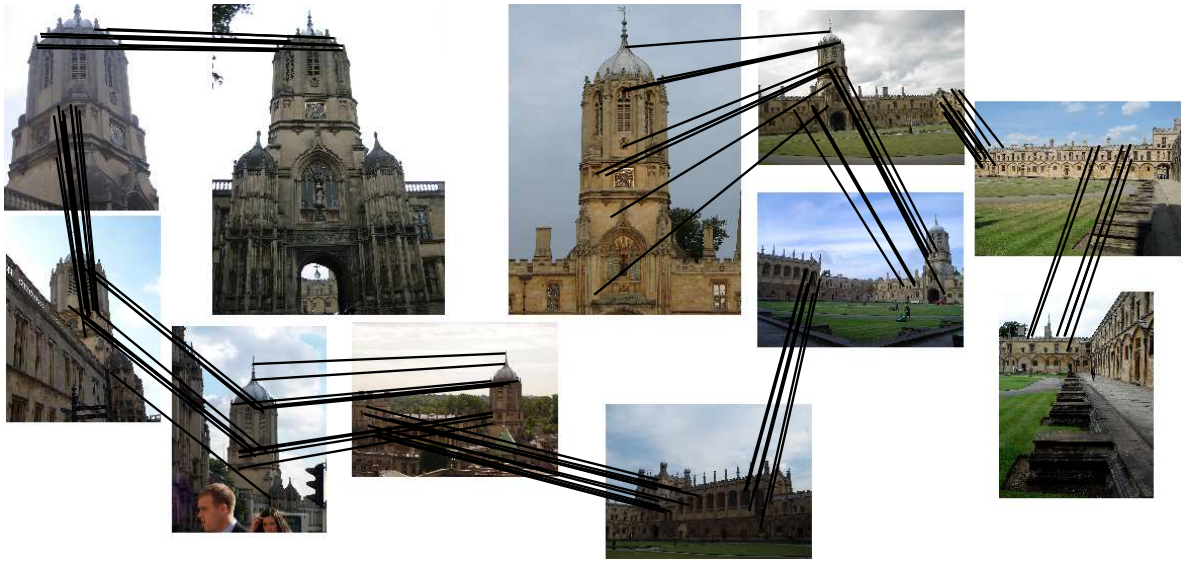


Fig. 1. Visualization of a part of a cluster of spatially related images automatically discovered from a database of 100K images. Only part of the cluster is shown. Overall, there are 113 images in the cluster, all correctly assigned. A sample of geometrically verified correspondences is depicted as links between the images. Note that the images show the tower from opposite sides.

The above mentioned process starts with an image provided or selected by the user. However, 3D registration is still a slow process. In general, it is not possible to do it online, and an immediate response to the user requires that 3D reconstruction is already available, computed off-line. The clustering method proposed in the paper is a suitable back-end for such a system, as it discovers sufficiently large sets of overlapping images suitable for automatic reconstruction. Moreover, it outputs inter-image correspondences that may bootstrap the 3D scene reconstruction process. Availability of sufficient number of images is essential for the 3D reconstruction, and almost all sets that are usable for 3D reconstruction have a size where our method retrieves the cluster almost certainly.

The rest of the paper is structured as follows. Section 2 reviews the work on unsupervised object and scene discovery, Section 3 describes the use of min-Hash for data mining purposes. In Section 4 the method is experimentally verified on real image databases.

2 Related work on unsupervised object and scene discovery

The problem of matching (organization) of an unordered image set was first addressed by Schaffalitzky and Zisserman in [10]. Their objective was first automatic recovery of geometric relations between images from a spatially related set (of tens of images) and then 3D reconstruction. We are interested in a similar problem, but also in discovery of multiple such sets in databases with several orders of magnitude higher number of images.

Recently, the majority of image retrieval systems adopt the bag-of-words approach [11], which we also follow. First, regions of interest are detected [12] and described by an invariant descriptor [13]. The descriptors are then vector quantized into a vocabulary of visual words [11, 5, 6].

The approach closest to ours is [14] by Sivic and Zisserman whose objective is unsupervised discovery of multiple instances of particular objects in feature films. Object hypotheses are instantiated on neighbourhoods centered around regions of interest. The neighbourhoods include a predefined number of other regions and the hypothesized object is represented by a fixed number of visual words describing the regions. Each hypothesized object is used as a query against the database consisting of key frames of the film. To reduce the number of similarity evaluations, which each requires counting the number of common visual words, only neighbourhoods centered at the same visual word are compared.

The method requires $\sum_{i=1}^{w} d_i^2$ similarity evaluations, where w is the size of vocabulary and d_i is the number of regions assigned to i-th visual word. Let D be the number of documents and t the average number of features in an

image, so that $\sum_{i=1}^{w} d_i = tD$. The lower bound on the complexity of the approach in [14] can be written as

$$\sum_{i=1}^{w} d_i^2 \geq \sum_{i=1}^{w} \left(\frac{tD}{w}\right)^2 = \frac{t^2}{w} D^2. \quad (1)$$

The asymptotical complexity of [14] is thus $\mathcal{O}(D^2)$. The factor t^2/w is a ratio of two constants independent of the size of the database. The size of the vocabulary commonly used is up to $w = 1,000,000$, the average number of regions in an image for the database used in this paper is slightly over $t = 2,800$, leaving the value of the coefficient $t^2/w = 7.84$ in order of units. Hence, the algorithm would behave as quadratic in the number of images even for relatively small databases. The complexity of [14] is thus the same as the complexity of querying the whole database with each image in turn. Such complexity is prohibitive for large databases.

Methods for query speed-up [15, 7] proceed by pre-clustering documents into similar groups. For a query, first a set of relevant document clusters is retrieved sub-linearly and the query is only evaluated against images in the selected clusters. Such an approach trades off recall for up to seven fold speed-up [7]. A speed-up of this order is insufficient to allow querying by each image on large databases.

Approaches improving the accuracy of image retrieval [7, 9] are relevant to this paper despite not helping seed initialization since they improve the second stage of our approach, the crawl for images visually connected to seed pairs. Accuracy improving techniques include learning a local inter-document distance measure based on the density in the document space [7] and selecting the most informative features for the vocabulary[9]. Note that the statistics used in those approaches might be difficult to update when new, either related (changing the density in the document space) or completely unrelated (changing the relevance of the features) images are inserted into the database.

Another class of methods tackling unsupervised object learning is based on topic discovery [16] via generative modeling like probabilistic Latent Semantic Analysis (pLSA) [17] and Latent Dirichlet Allocation (LDA) [18]. Object discovery based on topic analysis method was further developed in [19] where multiple segmentations were used to hypothesize the locations and extent of possible objects.

The pLSA and LDA models are a favourite choice for (unsupervised) object / image *category* recognition due to their generalization power. However, the ability to generalize to a topic such as "building" is rather a disadvantage when particular objects are sought.

We consider topic analysis approaches not suitable for our problem for the following reasons: (i) Speed: These learning methods are slow, iterative and sequential (difficult or impossible to parallelize). (ii) Topics discovered by pLSA / LDA typically appear in a number of images proportional to the size of the dataset while in this paper we aim at finding clusters of certain size independent of the size of the database. (iii) When new images are inserted into the database and a new topic should be formed using both old and new data, the methods need to process the original (already processed) data again together with the new ones.

3 Data Mining with min-Hash

In this section, the proposed method for discovery of clusters of spatially overlapping images is described. As the first step, pairs of images that are likely to be spatially overlapping, the so-called seeds, are found by a procedure exploiting properties of the min-Hash algorithm. Understanding the procedure requires at least a basic familiarity with min-Hash and we therefore review the algorithm in Sect. 3.1. Next, the four steps of the cluster discovery algorithm are detailed:

1. **Hashing.** Image descriptors are hashed into a hash table. In experiments in the paper we use 2^{51} different descriptor values. The probability of two images falling into the same bin (*exact* descriptor match) is proportional to their similarity – equation (2).
2. **Similarity estimation.** For all n-choose-2 pairs of the n images that have been hashed into the same bin a similarity is estimated. Similarity estimation is fast and consists of comparing two vectors and counting the number of identical elements. In this work, the number of vector elements is 512. The similarity is then thresholded.
3. **Spatial consistency.** For each image pair that passed the similarity test, spatial consistency is verified. Image pairs that pass spatial consistency test are the cluster seeds.
4. **Seed growing.** Once cluster seeds are generated, the seed images are used as visual queries and query expansion technique is used to 'crawl' the images in the cluster.

3.1 The min-Hash algorithm review

The min-Hash algorithm [20, 8] is a randomized method based on hashing finds highly similar image pairs with probability close to one, unrelated images with probability close to zero, and similar image pairs (with low but non-negligible similarity, such as images of the same object) with a rather small probability. The low recall stops the min-Hash from being used directly as a general image retrieval method. However, in this paper we argue that it can be efficiently used for data mining purposes.

A brief review of the min-Hash algorithm follows; for detailed description see [21, 20]. For the purpose of min-Hash, images are represented as sets of visual words. This is a weaker representation than a bag of visual words since the frequency is reduced into a binary information (present / absent). Similarity of two images $\text{sim}(A_1, A_2)$ is measured as a set overlap (ratio intersection over union) of their set representation

$$\text{sim}(\mathcal{A}_1, \mathcal{A}_2) = \frac{|\mathcal{A}_1 \cap \mathcal{A}_2|}{|\mathcal{A}_1 \cup \mathcal{A}_2|} \in \langle 0, 1 \rangle. \tag{2}$$

A min-Hash is a function f that assigns a number to each set of visual words (each image representation). The function has a property that the probability of two sets having the same value of the min-Hash function is equal to their similarity

$$P(f(\mathcal{A}_1) = f(\mathcal{A}_2)) = \text{sim}(\mathcal{A}_1, \mathcal{A}_2).$$

To estimate the similarity of two images, multiple independent min-Hash functions f_i are used. The fraction of the min-Hash functions that assigns an identical value to the two sets gives an unbiased estimate of the similarity of the two images.

Retrieval with min-Hash. So far, a method to estimate a similarity of two images was discussed. To efficiently retrieve images with high similarity, the values of min-Hash function f_i are grouped into s-tuples called sketches. Similar images have many values of the min-Hash function in common (from the definition of similarity) and hence have high probability of having the same sketches. On the other hand, dissimilar images have low chance of forming an identical sketch. Identical sketches are efficiently found by hashing.

The probability of two sets having at least one sketch out of k in common is

$$P(\text{collision}) = 1 - (1 - \text{sim}(\mathcal{A}_1, \mathcal{A}_2)^s)^k. \tag{3}$$

The probability depends on the similarity of the two images and on the two parameters: s the size of the sketch and k the number of (independent) sketches. These are the parameters of the method. Figure 2 visualizes the probability of collision plotted against the similarity of two images for fixed $s = 3$ and $k = 512$. Figure 3 shows different image pairs and their similarity.

3.2 Cluster seed generation

In this section, a randomized procedure that generates seeds from possible clusters of images is described. Let us first look at the plot of the probability of sketch collision against the similarity of the images depicted in figure 2. The sigmoid-like shape of the curve is important for the near duplicate detection task [20]. Image pairs with high similarity are retrieved with a probability close to one. Then, the probability drops rapidly - through similar image pairs (typically images of the same object from a slightly different viewpoint) that are occasionally retrieved to unrelated image pairs (with similarity below 1%) that have close to zero probability of being retrieved.

Now, for the purpose of data mining, we focus on the bottom left corner of the graph. According to equation (3) an image pair with similarity $\text{sim} = 0.05$ has probability 6.2% to be retrieved (using 512 sketches of size 3). Such a poor recall is certainly below acceptable level for a retrieval system. However, we aim at retrieving all relevant images from the image clusters in a single step. The task is to quickly retrieve seeds from the clusters - it is sufficient to retrieve a single seed per cluster, and we are fortunate that the importance of a cluster is related to its size in the database.

The probability that not a single image pair (seed) is found by the min-Hash depends on two factors - the similarity of the images in the cluster and the number of image pairs that actually observe the same object. In the following analysis, which demonstrates a lower bound on this probability, we assume that a particular object or landmark is seen in v views. We also assume that all the image pairs have the same (average) similarity ε. The probability that none of the pairs of v views is retrieved is

$$P(fail) = (1 - \varepsilon)^{\frac{v(v-1)}{2}}.$$

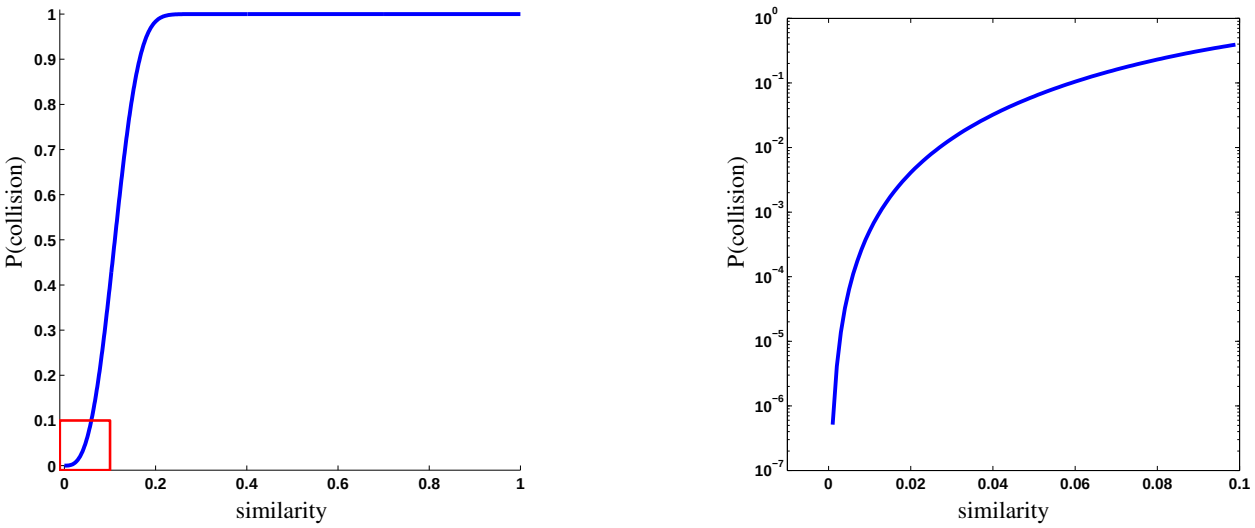


Fig. 2. The probability of at least one sketch collision for two documents plotted against their similarity; with $k = 512$ sketches, $s = 3$ min-Hashes per sketch. The right plot shows a close-up of the bottom left corner of the left plot. Note the logarithmic vertical axis.

Fig. 3. Examples of the relation of 'visual similarity' and the set overlap similarity.

The plot in figure 4 shows that for popular places (i.e. those where photos are often taken from) the probability of failure to retrieve an image pair vanishes. There are three plots for similarities 5%, 6% and 7%. Since the similarity is defined as a ratio of the size of the intersection over the size of the union, the difference between similarity 6% and 5% is substantial. Going from 6% to 5% similarity means removing 17.5% of elements that were in the intersection.

It is important to point out that the probability of finding a seed depends on the image similarities and the number of views and is completely *independent* of the size of the database. The v views have the same chance to be discovered in a database of 5000 images as in a database of several millions of images without any need to change the method parameters or re-hash. This is not true for many topic discovery approaches.

Time complexity. The method is based on hashing with a fixed number M of bins. The number of bins is based on the size of the vocabulary which cannot be infinitely increased without splitting descriptors of the same physical region. Assuming uniform distribution of the keys, the number C of keys that fall into the same bin is a random variable with a Poisson distribution where the expected number of occurrences is $\lambda = D/M$. The expected number of key pairs that fall into the same bin (summed over all bins) is

$$\sum_{i=1}^{M} \mathbf{E}(C^2) = \sum_{i=1}^{M} \left(\lambda^2 + \lambda\right) = \frac{D^2}{M} + D. \tag{4}$$

The asymptotical time complexity is $\mathcal{O}(D^2)$ for D, i.e. size of the image database, approaching the infinity. However, for finite databases of sizes up to $D \leq M$, the method behaves as linear in the number of documents since $D^2/M +$

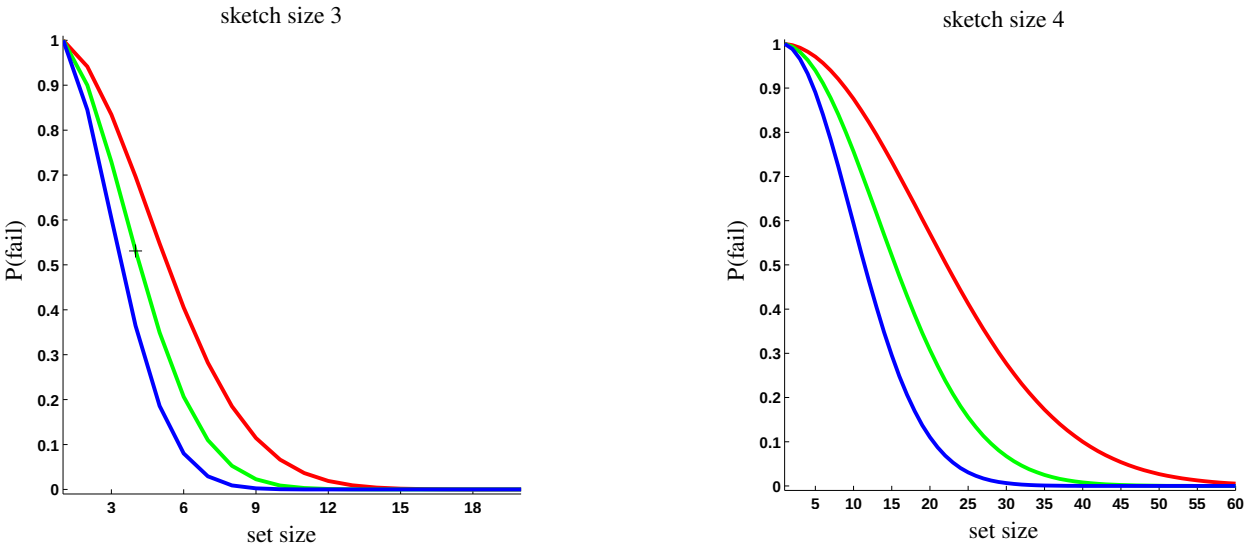


Fig. 4. Probability of failure to generate a seed in a set of images depicting the same object using min-Hash with 512 sketches of size 3 (left) and 4 (right); note the different scales on the horizontal axes. The three curves show the dependence for different 'average' similarity equal to 7% (lowest curve), 6% (middle) and 5% (highest). The marker '+' on the left plot denotes experimental result on the University of Kentucky dataset (cluster size 4, P(fail) = 1 - 0.469), see section 4.1.

$D \leq 2D$. In the min-Hash algorithm, the number of keys depends on the size of the vocabulary w and the size of the sketch s and is proportional to $M = w^s$. In the experiments in this paper, we used $w = 2^{17}$ and $s = 3$ or $s = 4$. This gives the number of different hash keys $M = 2^{51}$ and $M = 2^{68}$. We believe that this number is sufficient to conveniently deal with web scale databases.

3.3 Growing the seed

We build on the query expansion technique [8] to increase the recall. The idea is as follows: an original query is issued and the results are then used to issue new query. Not all results are used, only those that have the same spatial feature layout (for more details on spatial verification see the following section). The spatial verification prevents the query expansion from so-called topic drift, where an unrelated image is used to expand the query.

In our approach, we combine two types of query expansion methods suggested in [8] - transitive closure and average expansion. In the transitive closure, each previously unseen (spatially verified) result is used to issue a new query. This method is used to 'crawl' the scene. To improve the recall, each query is attempted to be expanded by an average expansion: Result images in which a sufficient number of regions are related by a homography (homographies) to the query image are selected. The homography is then used to back-project features from the result image to the query image (only features within a bounding box of the homography support are mapped). A new query is issued using the combination of the original features and the features gathered from the result images. For efficiency, each image is used at most once for an average query expansion.

If our data mining method is used for obtaining images for 3D reconstruction, a (partial) 3D model can be used for query expansion [8]. To retrieve images from unexplored viewpoints synthetic views (not necessarily pixel-wise) could be generated and used as queries. This is beyond the of scope of this paper.

3.4 Spatial verification

In spatial verification we build on the many-to-many RANSAC-like approach from [6]. Tentative correspondences are defined by a common visual word ids. The geometric constraint is an affine transformation. This choice is convenient since a single ellipse-to-ellipse correspondence (plus constraint on the gravity vector) is sufficient to instantiate the model. The model of affine transformation with loose thresholds allows for detection of close-to-planar structures in the scene with no significant perspective distortion. Unlike in [6], we fit multiple such models. The global consistency of those models is then verified by a RANSAC fitting of an epipolar geometry or homography [22]. This final check is rapid – tentative correspondences for this stage are a union of inlier correspondences from the previous stage and a

high inlier ratio is expected (only a few samples are drawn in RANSAC). Since we are fitting an exact model now, the geometric thresholds are set tight.

There are two common sources of mismatches: degenerate configurations of points (close to collinear point sets) and repeated structure (many features assigned to a single visual word, typically repeated in a grid-like structure). In our implementation, in order to positively verify a pair of images there has to be a sufficient number of matches that are not part of a degenerate or repeated structure.

4 Experimental Evaluation

We have conducted two experiments. The first one checks whether the probability of seed generation is sufficiently high on real data as predicted by theoretical estimates presented in section 3.2. In the second experiment, clusters of spatially related images are discovered in a database of 100K images.

4.1 Seed generation success rate

To evaluate the success rate of the seed generation stage on real data, we use a standard image retrieval benchmark dataset - the University of Kentucky dataset, introduced in [5]. This database contains 10200 images; a group of 4 images depicts the same object / scene, i.e. there are 2550 clusters of size four. The standard experiment on the database is to query the database with each image in turn trying to retrieve the other three images from the cluster. Success of the retrieval is measured by the average number of correctly returned images in the top four results (the query image is to be retrieved too). The perfect score is thus 4.

We, however, are interested in a different statistic. Our objective is to measure for how many clusters (all of size four) the proposed method generates at least one seed. For this experiment, we have used a visual vocabulary of 2^{17} visual words. For each image, 512 independent random min-Hash functions were evaluated and grouped into 512 sketches of size 3 (individual min-Hashes were used multiple times). With this setting, there are 11556 pairs of images with at least one common sketch value (a sketch collision) of which 3553 passed the similarity test at 0.045 (step 2 of the clustering procedure); out of the 3553 seeds 3210 were within a ground-truth defined group of four images. The number of clusters of four images for which at least one pair was suggested by the hashing is 1196 (out of 2550 possible clusters). In other words, a seed for a cluster of size four is generated with a probability of 46.9%, which is very close to the expected value of failure, see figure 4, left plot.

Note that in this experiment we are interested only in the false negative rate of the seeding process, not the false positive rate. Potential seeds that are not within a group of four ground truth images are not necessarily false positives as many objects are presented on the same background. According to the ground truth for the database, such images are in different groups, i.e. spatially unrelated, despite having a significant spatial overlap on the background.

We did not perform spatial verification filtering stage on this dataset. We used only data provided by the authors of the database which do not include information necessary for geometry verification.

We are aware that using the estimated similarity to order the results for each query image, the standard retrieval score for this method is 1.63. Such a score does not compete with current retrieval systems with scores between 3.3 and 3.6. Yet, the min-Hash method proves suitable for randomized data mining by seed generation.

4.2 Clustering on the 100K Oxford Landmark Database

We conducted an experiment on a large database of images downloaded from Flickr [3]. This database contains $5,062$ images from publicly available Oxford Landmark Database [23] and $99,782$ from *Flickr1* dataset[3] used in [6]. Both sets are composed of high resolution images (1024×768). Together, there are $104,844$ images with $294,105,803$ features (2805 features per image on average). The SIFT descriptors of the features take 35GB. In this dataset, 11 landmarks were manually labelled in the images. There are four possible labels for an image (i) Good - a nice, clear picture of the object, (ii) OK more than 25% of the object is clearly visible, (iii) Junk less than 25% of the object is visible, or there is a very high level of occlusion or distortion, and (iv) Absent the object is not present.

As in the previous experiment, we used a vocabulary of 2^{17} visual words for the min-Hash seed generation; a 1M vocabulary was used for seed growing. The database contains clusters of size of tens (up to hundreds) of images. To show the potential of the method, we used 512 min-Hashes grouped into 512 sketches of size three. These settings allow to discover even small clusters of several images with reasonable probability and are the same as in the University

[3] Courtesy of VGG, University of Oxford

	Good	OK	sketch 3 (%)	unrelated	sketch 4 (%)
All Souls	24	54	97.44	0	97.44
Ashmolean	12	13	68.00	0	0
Balliol	5	7	33.33	0	0
Bodleian	13	11	95.83	1	95.83
Christ Church	51	27	89.74	0	89.74
Cornmarket	5	4	66.67	0	0
Hertford	35	19	96.30	1	0
Keble	6	1	85.71	0	0
Magdalen	13	41	5.56	0	1.85
Pitt Rivers	3	3	100.00	0	0
Radcliffe Camera	105	116	98.64	0	98.46

Table 1. Results for annotated images in the Oxford Building Dataset. The first two columns show the number of ground truth images labelled 'Good' and 'OK' respectively. The column 'sketch 3' displays the percentage of labelled images that were clustered into a single cluster using min-Hash with sketches of size three, 'unrelated' gives an absolute number of unrelated images in that cluster. The column 'sketch 4' presents results for sketches of size four.

of Kentucky database experiment. On average, the min-Hash generated 38.4 sketch collisions per image. These were reduced to 1.23 potential seeds per image by thresholding the estimated similarity at 0.045 - this corresponds to $129,341$ protectional seeds. Out of those, 3103 images were found to have an exact duplicate in the database (the same image was downloaded under different user tags), and 289 images were found to have a near duplicate (see figure 6). Both exact and near duplicates were dropped and the remaining potential seeds were subject to spatial verification, leaving 441 verified seeds. This number is an upper bound on the number of clusters, since typically there are multiple seeds per cluster. The seed growing by query expansion discovered 354 distinct clusters covering $2,643$ images. Cluster examples are shown in figure 5 and also in figure 1.

Table 1 summarizes the results on objects with ground truth information. For each landmark, we found cluster containing the most positive (Good and OK) images of that landmark and computed the fraction of positive ground truth images in this cluster. Also, the absolute number of unrelated images is reported by eye-balling these clusters. Other buildings that appear in the same cluster are not considered unrelated as long as there exist images that link these objects. For example, images of All Souls and the Radcliffe Camera are all in one cluster - they are right next to each other and even appear together on several images.

Clusters corresponding to all ground-truth objects were successfully discovered with the exception of the 'Magdalen' tower. The percentage of images assigned to the relevant cluster is consistent with the retrieval results in [6, 8] and is related to the 'difficulty' of each landmark. This also holds for the 'Magdalen' - retrieval results were by far the worst for this landmark. In our case, three images of the tower were discovered and the method was unable to spatially verify and grow to any other image.

The choice of the size of sketch equal to three is suitable for demonstrating the power of the method on a not-so-large database of 100K images. It retrieves even small, perhaps even uninterestingly small, clusters. These settings will not be acceptable for web scale database size of more 10^7 images or more. To simulate real conditions, we have also used 512 sketches of size four, which is suitable for very large databases, but returns with acceptable probability only larger clusters. Still, the size of discovered clusters is comparable (smaller) than the size of clusters used in Photo Tourism [4]. The four largest clusters from the Oxford Landmark ground truth were discovered (together with other larger clusters that are not included in the ground truth). In the case of Magdalen tower, it is seen on one image of different cluster (figure 5 second cluster).

Timing. The seed generation took 26 min 39 sec and the seed growing took 16 min 20 sec on a PC using a single processor (MATLAB / MEX implementation). The complete processing of the database took thus less than 45 minutes, which is 0.025 seconds per processed image. Note that all steps of the proposed method are easy to parallelize.

5 Applications

We show two potential applications that are made possible by cluster discovery – landmark discovery/labelling and automatic 3D reconstruction.

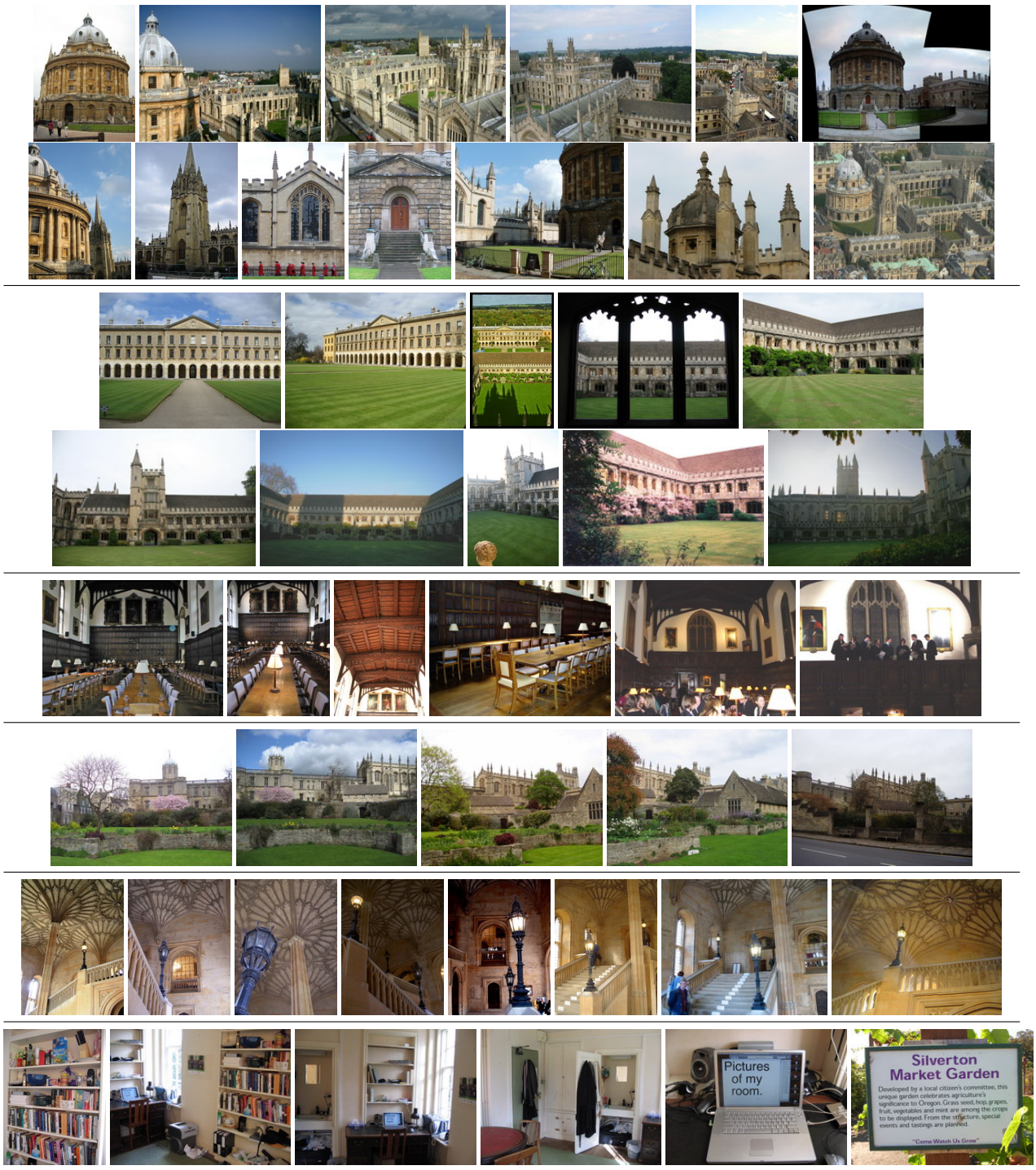


Fig. 5. Selected images from selected clusters discovered in the 100K database including the Oxford Landmark dataset. Top - the largest cluster containing the Radcliffe Camera and All Souls (404 images). Below - discovered clusters of sizes 53, 14, 51, 18, and 13 respectively, not in the ground truth annotation. The last cluster contains one false positive (the rightmost image), the other clusters are visually connected. The top four clusters were also discovered in the experiment with sketches of size four.

Fig. 6. A useful side-effect: a sample of near duplicate images detected in the database.

Fig. 7. Images with two different object segmented. Radcliffe Camera and All Souls (left), Bridge of Sighs and Bodleian Library (right). The extent of the landmarks is shown in different markers and colours.

Landmark labelling. As mentioned before, a single cluster may contain more than one landmark. We can further factorize each cluster (using image matches and weak 3D constrains such as co-planarity and disparity) to sub-clusters containing a single landmark. Results of automatic landmark segmentations are shown in figure 7. Matches between landmark sub-clusters are shown in figure 8. Finally, common user annotations from the sub-clusters (if available) may serve as name tags as shown in figure 9. The segments and the positions for the labels were discovered automatically, the correct textual annotations were added manually.

Full 3D Reconstruction. The discovered clusters were processed by a 3D reconstruction pipeline [24]. Sample results are shown in figures 10 and 11.

6 Conclusions

We have proposed a method for discovering spatially-related images in large scale image databases. Its speed depends on the size of the database and is close to linear for database sizes up to approximately $2^{34} \approx 10^{10}$ images. The success rate of cluster discovery is dependent on the cluster size and average similarity and is independent of the size of the database. The properties and performance of the algorithm were demonstrated on datasets with 10^4 and 10^5 images. The proposed algorithm provides, as a side effect, a stat-of-the-art near duplicate image detection.

The desirable characteristics of the data mining method stem from an appropriate use of the min-Hash algorithm, which has been so far used only for near duplicate (high similarity) problems.

Acknowledgments. Authors would like to thank to Daniel Marincec for the 3D reconstruction, Michal Perďoch for discussions and help, and James Philbin for providing the data and his implementation of the spatial verification [6]. We are grateful for support from EC grant 215078 DIPLECS.

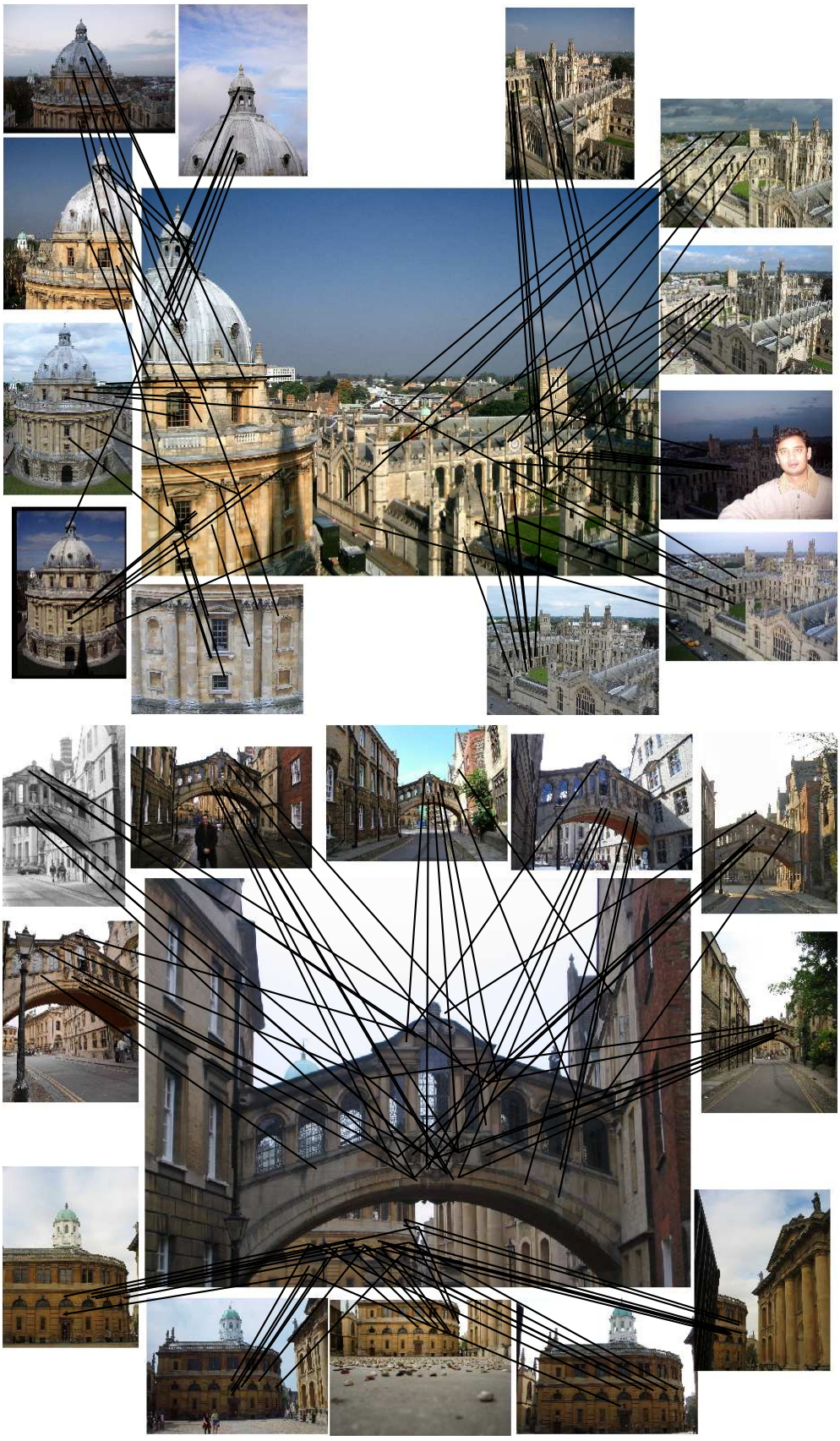


Fig. 8. Automatic localization of different buildings - matching between image groups. Only a small fraction of matches is shown.

Fig. 9. Automatic labelling of objects.

References

1. http://books.google.com/help/maps/streetview/ (www)
2. http://www.panoramio.com/ (www)
3. http://www.flickr.com/ (www)
4. Snavely, N., Seitz, S., Szeliski, R.: Photo Tourism: exploring photo collections in 3D. In: Proc. ACM SIGGRAPH. (2006) 835–846
5. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: Proc. CVPR. (2006)
6. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. CVPR. (2007)
7. Jegou, H., Harzallah, H., Schmid, C.: A contextual dissimilarity measure for accurate and efficient image search. In: Proc. CVPR. (2007)
8. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: Proc. ICCV. (2007)
9. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: Proc. CVPR. (2007)
10. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In: Proc. ECCV. Volume 1., Springer-Verlag (2002) 414–431
11. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proc. ICCV. (2003)
12. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. IJCV **65** (2005) 43–72
13. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60** (2004) 91–110
14. Sivic, J., Zisserman, A.: Video data mining using configurations of viewpoint invariant regions. In: Proc. CVPR. (2004)
15. Fraundorfer, F., Stewénius, H., Nistér, D.: A binning scheme for fast hard drive based image search. In: Proc. CVPR. (2007)
16. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collections. Technical Report A. I. Memo 2005-005, Massachusetts Institute of Technology (2005)
17. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR. (1999)
18. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. J. Machine Learning Research **3** (2003) 993–1022
19. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2006)
20. Chum, O., Philbin, J., Isard, M., Zisserman, A.: Scalable near identical image and shot detection. In: Proc. CIVR. (2007)
21. Broder, A.: On the resemblance and containment of documents. In: SEQS: Sequences '91. (1998)
22. Frahm, J.M., Pollefeys, M.: RANSAC for (Quasi-)Degenereate data (QDEGSAC). In: Proc. CVPR. (2006)
23. http://www.robots.ox.ac.uk/∼vgg/data/oxbuildings/ (www)
24. Martinec, D., Pajdla, T.: Robust rotation and translation estimation. In: CVPR. (2007) 8

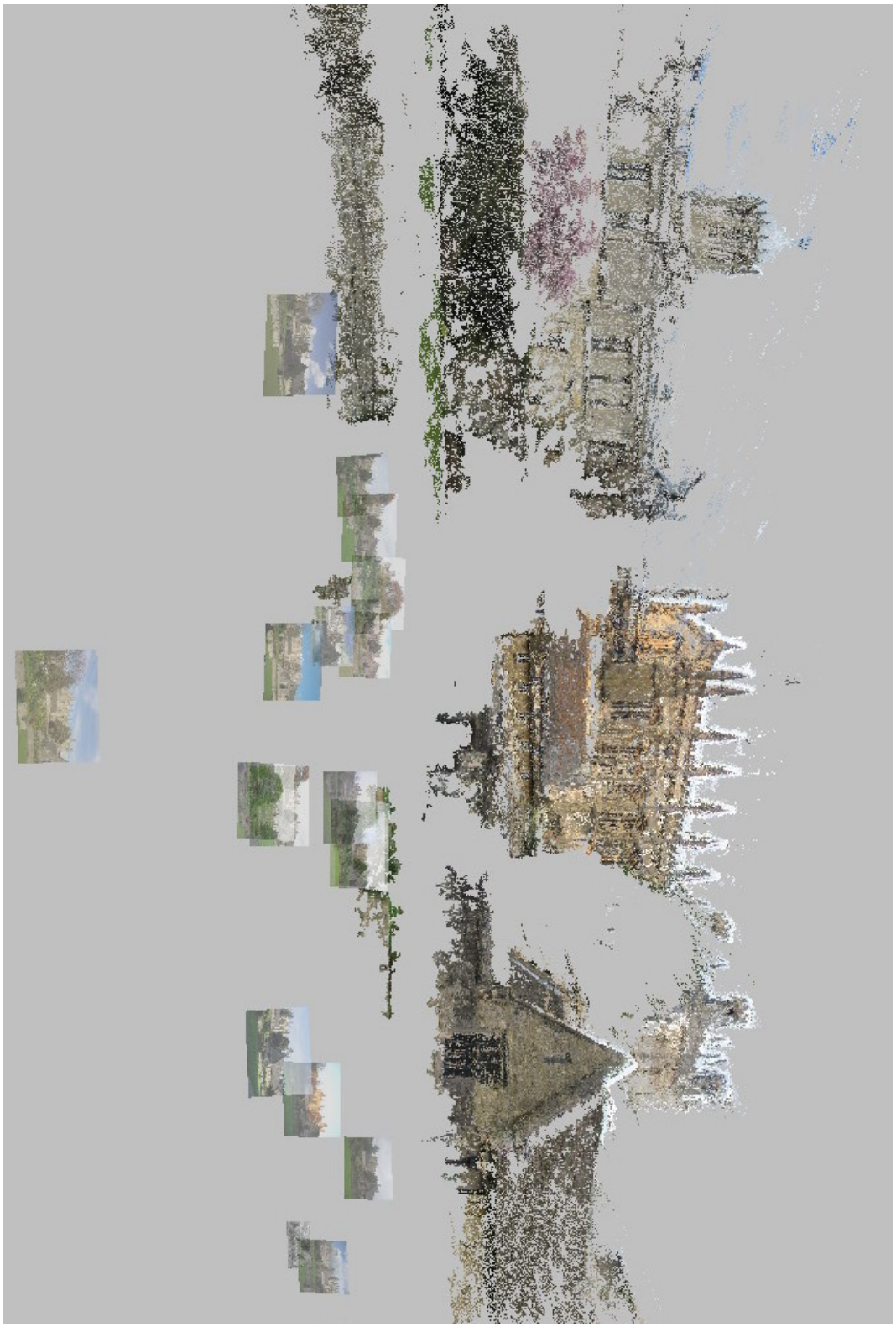


Fig. 10. Automatic 3D reconstruction form a discovered cluster.

Fig. 11. The 3D scene visible in the top image from the viewpoint of a different image.