Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Visual synonyms for landmark image retrieval

Efstratios Gavves *, Cees G.M. Snoek, Arnold W.M. Smeulders

Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 13 December 2010
Accepted 16 October 2011
Available online 7 November 2011

Keywords:
Image representation
Image retrieval

ABSTRACT

In this paper, we address the incoherence problem of the visual words in bag-of-words vocabularies. Different from existing work, which assigns words based on closeness in descriptor space, we focus on identifying pairs of independent, distant words – the visual synonyms – that are likely to host image patches of similar visual reality. We focus on landmark images, where the image geometry guides the detection of synonym pairs. Image geometry is used to find those image features that lie in the nearly identical physical location, yet are assigned to different words of the visual vocabulary. Defined in this way, we evaluate the validity of visual synonyms. We also examine the closeness of synonyms in the $L2$-normalized feature space. We show that visual synonyms may successfully be used for vocabulary reduction. Furthermore, we show that combining the reduced visual vocabularies with synonym augmentation, we perform on par with the state-of-the-art bag-of-words approach, while having a 98% smaller vocabulary.

1. Introduction

In recent years several local visual features have been proposed, which encode the richness of localized visual patches [1,2]. Although these features perform well in object and concept recognition as exemplified in the advances of TRECVID and PASCAL [3,4], the detection and transformation of the visual reality of an image patch into a feature vector is far from perfect [5,6]. Despite this fact and to the best of our knowledge, there has been so far limited research of the high dimensional visual feature space formed and its properties.

For their ability to capture local visual information well enough, local feature detectors and descriptors are mostly used. Feature detectors and descriptors operate directly on the raw visual data of image patches, which are affected by common image deformations. These image deformations affect either image appearance, which accounts for the way the image content is displayed, or image geometry, which accounts for the spatial distribution of the image content inside the image. Image appearance variations include the abrupt changes of illumination, shading and color constancy [7]. Image geometry variations are related to viewpoint changes, non-linear scale variations and occlusion [8–12]. Several feature descriptors that provide invariance against image appearance deformations have been proposed [7]. However, there are no specific features that deal adequately with image geometry deformations. Instead, this level of invariance is partly reached on the next level of image representation, using for example the bag-of-words model [13–16]. Despite this *a posteriori* acquired invariance under geometric deformations, feature vectors of similar visual reality are still erroneously placed in very different parts of the feature space. Thus, the image feature space spanned by local feature detectors and descriptors is fuzzily populated.

Moreover, to be sufficiently rich to capture any local concept the visual feature space has to be of high dimensionality. However, distance measures in high dimensional spaces exhibit a more sensitive nature [17]. Thus distance measures, a cornerstone of most machine learning algorithms, are less indicative of the true similarity of two vectors, which as a result disturbs the image retrieval process. Therefore, error-prone distance measures also contribute to the fuzzily populated feature space.

By treating local image descriptors as orderless words, images in the bag-of-words model may be classified in a class on the basis of word histograms. In effect, bag-of-words hopes for large number statistics to even out the consequences of the aforementioned image deformations. Words are obtained by clustering in the descriptor space [18], implicitly assuming that all patches covered by one word represent the same part of reality. And, that different clusters correspond to different parts of reality. These clusters lie inside the fuzzily populated feature space, resulting in visual words that have little coherence in the semantics of the patches they contain, see Fig. 1. For standard challenges, like PASCAL which targets at general object recognition, visual word incoherence does not affect the performance drastically and vocabularies of size up to 4 K clusters suffice. However, for more challenging datasets, like *Oxford5k* [14] or [19], image appearance and geometry deformations start to have a much greater impact. Hence techniques that make better use of the feature space are needed. For complex datasets, larger vocabularies have proven to operate more effectively [14,19].

[*] Corresponding author.
E-mail address: egavves@uva.nl (E. Gavves).

1077-3142/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2011.10.004

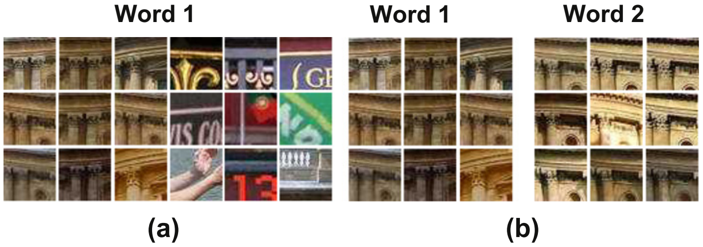


Fig. 1. (a) Image patches mapped to one visual word of the bag-of-words vocabulary. Note the visual incoherence. (b) Comparison between image patches from two different words. Note their perceptual similarity.

Larger vocabularies fragment feature space finer yielding visual words that are more concise, albeit less populated. Despite their effectiveness, large vocabularies merely postpone rather than solve the problem of the fuzzily populated feature space. Another technique that helps to ameliorate the errors during feature acquisition is the use of soft assignment for mapping features to clusters. Unlike hard assignment that performs a crude binary assignment to a single cluster, soft assignment distributes the probability mass of the mapping to a number of adjacent clusters [20]. Unfortunately, soft assignment compensates only for the assignment errors near the cluster borders. Errors that might occur because of the misplacement of features in distant parts of the feature space remain unaffected.

In this paper we propose visual synonyms, a method for linking semantically similar words in a visual vocabulary, let them be distant in feature space or not. The bag-of-words model is used on landmark images, because their unchanged geometry allows for mapping between different images with different recording conditions, which opens the door to perspectives for linking words as synonyms. When a link to the same spot is found, it is clear the word represents nearly the identical patch in reality. However, due to the accidental recording conditions in each of the words, the features may differ significantly. Thus, this link establishes a connection between two parts of the feature space, which, despite their distance, correspond to image patches of similar visual reality. Visual synonyms comprise a vehicle for finding the parts of feature space, which are nearly identical in reality. This allows for further refinement of visual word definitions. Also, visual synonyms can be used for vocabulary reduction. By using a fraction of visual synonym words, we are able to reduce vastly the vocabulary size without a prohibitive drop in performance.

This paper extends [21] with additional experiments and a more deep analysis of the behavior of visual synonyms and visual words. The rest of the paper is organized as follows. In Section 2 we present some related work. In Section 3 we introduce the notion of visual synonyms and we propose an algorithm for their extraction. We describe our experiments in Section 4 and we present the results in Section 5. We conclude this paper with a short discussion of the acquired results.

2. Related work

The bag-of-words method is the state-of-the-art approach in landmark image retrieval [14]. The core element of the bag-of-words model is the vocabulary $W = \{w^1,\ldots,w^K\}$, which is a set of vectors that span a basis on the feature vector space. Given the vocabulary and a descriptor d, an assignment $q^r \in 1,\ldots,K$ to the closest visual word is obtained. We may construct the vocabulary W on a variety of ways, the most popular being k-means [22]. Based on the bag-of-words model, an image is represented by a histogram, with as many bins as the words in the vocabulary. The word bins are populated according to the appearance of the respective visual word in the image. Therefore, an image I is represented by $h_I = g(w_I^1),\ldots,g(w_I^K)$, where $g(\cdot)$ is a response function assigning a value usually according to the frequency of the visual word in the image. More advanced techniques have recently been proposed, better encoding the original descriptor d using the vocabulary basis W, thus yielding significant performance improvements, often at the expense of a high memory and computational cost [23] After obtaining the histogram of responses, all spatial information is lost. Following [14,24], we enrich the bag-of-words model with spatial information using homography mappings that geometrically connect pairs of images.

The most popular choice for feature extraction in the bag-of-words model is the SIFT descriptor [1]. Given a frame, usually split into a 4 × 4 grid, the SIFT descriptor calculates the edge gradient in eight orientations for each of the tiles in the grid. Thus resulting in a 128-D vector. Although originally proposed for matching purposes, the SIFT descriptor also dominates in image classification and retrieval. Close to SIFT follows the SURF descriptor [2]. SURF is designed to maintain the most important properties of SIFT, that is extracting edge gradients in a grid, while being significantly faster to compute due to the internal use of haar features and integral images.

An efficient and inexpensive extension to the bag-of-words model is visual augmentation [24,25]. According to visual augmentation, the retrieval of similar images is performed in three steps. In the first step the closest images are retrieved, using the bag-of-words model. In the second step, the top ranked images are verified. In the third step, the geometrically verified images lend their features to update the bag-of-words histogram of the query image and the new histogram is again used to retrieve the closest images. In the simplest case, the update in the second step averages over all verified images closest to the query [25,24]. In a more complicated scenario, the histogram update is based on a multi-resolution analysis of feature occurrences across various scenes [24]. For visual augmentation to be effective, the query's closest neighbor images have to be similar to the query image. Therefore geometric verification is applied. As expected, the top ranked images are usually very similar to the query image. However similar, these images exhibit slight differences due to their respective unique imaging conditions. These slight differences supplement the representation of the original query with the additional information that stems from the possible variations of visual reality as depicted in the image. Finally, the augmented query representation leads to a boost in performance. In this paper we draw inspiration from Chum et al. [24] and Turcot and Lowe [25], however we do not use any graph structure to connect images together.

Apart from image appearance, landmark scenes are also characterized by their unchanged geometry. Given that in a pair of images geometry changes because of translation, rotation and scale, there is a matrix that connects these two images together.

This matrix can be either the homography matrix or the fundamental matrix, according to the assumed geometry between the pictures, and can be computed using a robust iterative estimator, like RANSAC [26]. Faster RANSAC-based algorithms take advantage of the shape of local features to significantly speed up homography estimation [14]. Besides RANSAC-based geometry estimation, there are also other, less strict, techniques for taking geometry into account. Jegou et al. [27] check the consistency of the angle and scale value distribution of the features of two images. The rationale is that features extracted from the same physical location should have comparable scale and angle ranges. Pairs of images that exhibit dissimilar scale and angle distributions are considered as geometrically inconsistent. In [19] Wu et al. compare the spatial distribution of matched features between a pair of images. The motivation is that the feature distribution over one image should be as similar as possible with the one of another image. In the current work we follow the approach proposed in [14] to efficiently estimate pair-wise homography mappings.

In image retrieval and classification, current vocabulary sizes range from small, typically 4K [15] to large 1M words [14,19]. Because of the computational and storage requirements, large vocabularies are difficult to manage in real world scenarios that involve very large datasets. Therefore, methods for vocabulary reduction have been proposed. These methods try to keep retrieval or classification performance constant, whilst reducing the vocabulary significantly. Schindler et al. [28] and Zhang et al. [16] propose to retain words that appear to be frequent given the concepts in a classification task. In contrast, Turcot and Lowe [25] use geometrically verified visual words, which are appropriate for constructing a reduced vocabulary. First, they compare pairs of images geometrically with the use of RANSAC. Visual words returned as inliers from RANSAC are considered to be particularly insensitive to common image deformations. The vocabulary obtained is noticeably smaller without a significant loss in performance. However, for this technique the reduced vocabulary size is a dependent variable rather than an independent one. The size of the new vocabulary is not user-defined and depends on the geometric properties of the dataset pictures. In order to loosen this harsh geometric constraint, we propose a controllable selection of words from a pool of visual words, which are robust against common image deformations. Furthermore, in [25] visual words are found, which repeatedly appear in pictures of the same scene and are also geometrically consistent. Thus, variations of the visual elements in the very same scenes might be omitted, variations that are possibly important for recognizing the scene. We therefore investigate, as part of our approach, whether these variations should be taken into account in the vocabulary reduction process.

3. Visual synonyms

We define visual synonym words as *visual word pairs, which refer to image patches with similar visual appearance*. Similar visual appearance is common in images that depict the same, identical object of interest, like a famous building or monument. Examples of such images, which we refer to as landmark images, are "Eiffel tower, Paris" or the "All souls College, Oxford" pictures. Consequently, non-landmark images depict arbitrary objects, such as random people, and a random car. For landmark images visual synonym words cover descriptors that correspond to image patches originating from nearly identical physical elements.

To obtain visual synonyms we must find different visual words that are likely to correspond to visually similar patches. We need an independent information source to supply us with additional knowledge on the image's visual reality. Geometry is an independent information source, since it supplies information about the spatial properties of the image content. However, we lack the tools to confidently extract valid geometric information in object or abstract scene images. For this reason we opt for landmark images containing pictures of the same physical locations, whose largely unchanged geometry is ideal for geometry exploitation. Although other information sources may as well be used to analyze an image's visual reality, the proposed algorithm makes use of strict geometry, in effect optimizing for landmark image retrieval.

3.1. Preliminaries

We first introduce some notation. Following the query-by-example paradigm, we refer to a query image of dataset I as I_Q generating a ranked list I^j_Q, where j denotes the rank. We define image feature ξ as the local descriptor extracted on an interest keypoint with scale and location $\mathcal{X}$, mapped to a visual word w^r of the vocabulary. Consequently, the ith feature of image I_1 is denoted as $\xi_{1,i} = \{w^r_{1,i}, \mathcal{X}_{1,i}\}$. Finally, the homography matrix that relates the geometries of a pair of images I_Q and I^j_Q is denoted by $H(I_Q, I^j_Q)$.

3.2. Connecting visual words with geometry

Matching the geometry between two images is extensively studied in the literature [24,14,25,29]. The most prominent approach starts with a specific geometry assumption, like projective or epipolar geometry [26]. For projective geometry, the homography matrix H maps every point from the first image to one and only one point in the other image. Given only a few feature correspondences between a pair of images, we can calculate a candidate homography matrix that relates the images in a geometric fashion. Normally an iterative algorithm for generating hypotheses is used and the homography matrix that scores best is kept. Alternatively, the homography matrix can be calculated by following a deterministic approach without the use of any iterative scheme. It was shown in [14] that this approach leads to a significant computational speed-up. Homography matrix estimation, and image geometry in general, has mostly been used to reject images that have similar bag-of-words appearance but poor geometric consistency. Apart from this post-processing usage, geometry has not been exploited much for the visual representation of landmark images. In the current work we calculate the homography matrix for geometry estimation, in order to recognize the geometrical properties of the scene. We exploit geometrical scene properties to derive more effective representations for landmark appearance.

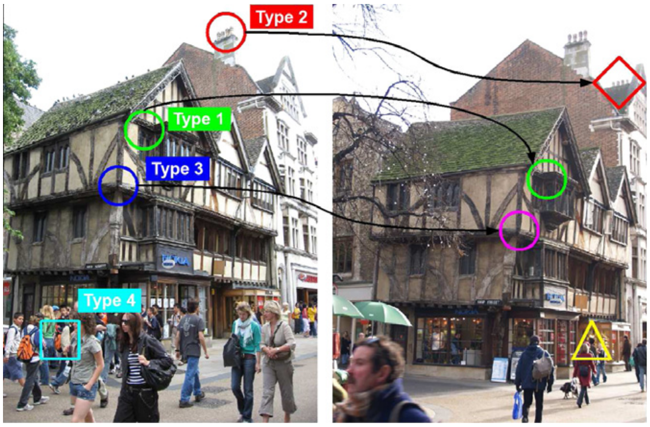


Fig. 2. Four different types of relation among words pairs. The same shape of the indicator (○, ◇) refers to same location, whereas same color refers to the same word. We are interested in pairs of words that have different visual appearance but refer to the same location in the world, that is pairs of words represented by the same shape but with different color (*Type* 3).

When two images I_1 and I_2 are connected with a matrix H, which is estimated using RANSAC, four possible feature pair relations exist between point pairs in the two images, see also Fig. 2.

Type 1: features ξ in nearly identical physical locations mapped to the same visual word w. That is

$$\xi_{1,i}, \xi_{2,j} : w_{1,i} = w_{2,j}, \mathcal{X}_{1,i} \approx H(I_1, I_2) \cdot \mathcal{X}_{2,j}.$$

We call these pairs visual metonyms.

Type 2: features ξ in different physical locations mapped to the same visual word w. That is

$$\xi_{1,i}, \xi_{2,j} : w_{1,i} = w_{2,j}, \mathcal{X}_{1,i} \neq H(I_1, I_2) \cdot \mathcal{X}_{2,j}.$$

We call these pairs visual homonyms.

Type 3: features ξ in nearly identical physical locations mapped to different visual words w. That is

$$\xi_{1,i}, \xi_{2,j} : w_{1,i} \neq w_{2,j}, \mathcal{X}_{1,i} \approx H(I_1, I_2) \cdot \mathcal{X}_{2,j}.$$

We call these pairs visual synonyms.

Type 4: features ξ in different physical locations mapped to different visual words w. That is

$$\xi_{1,i}, \xi_{2,j} : w_{1,i} \neq w_{2,j}, \mathcal{X}_{1,i} \neq H(I_1, I_2) \cdot \mathcal{X}_{2,j}.$$

These are random visual word pairs.

We consider a location in the world as nearly identical, when $|\mathcal{X}_{1,i} - H(I_1, I_2) \cdot \mathcal{X}_{2,j}| < \epsilon$.

Feature pairs of *Type* 1 and *Type* 2 are widely used in the literature as input to RANSAC [14,24,25]. Naturally, feature pairs of *Type* 4 make less sense to use in matching, whilst feature pairs of *Type* 3 have been ignored in the literature. However, feature pairs of *Type* 3 allow us to associate independent visual words of the vocabulary, which emerge from the same physical structure. This association provides us with the opportunity to find clusters in the descriptor space that have truly similar visual reality. This is a novelty with respect to state-of-the-art image retrieval and classification techniques [14,24,15]. Visual metonyms refer to visually similar patches that generate feature values being clustered to the same visual word, whereas visual synonyms refer to visually similar patches clustered to different visual words. Ideally, we would like to transform visual synonyms into metonyms. Since metonyms already exhibit a desirable and consistent performance, we focus on investigating the nature of visual synonyms.

3.3. Visual synonyms extraction

Our visual synonym extraction algorithm is a four-step procedure. During this algorithm, we are looking for potential visual synonym pairs, that is pairs of *Type* 3, see Fig. 2. For the extraction of visual synonyms a visual appearance similarity measure $d(\cdot)$ and a geometric similarity measure $g(\cdot)$ are used. We also use a threshold γ for assessing the geometric similarity of a pair of images, where γ refers to the minimum required number of inliers returned from RANSAC.

Step 1: Visual ranking. We rank all images in a data set according to their similarity with respect to a query image I_Q, using the standard bag-of-words model for modeling visual appearance. After this step, we obtain an ordered list $\{I_Q, I^j_Q\}$, such that:

$$d\left(I_Q, I^j_Q\right) < d\left(I_Q, I^{j+1}_Q\right), \; j = 1, \ldots, |I| - 1, \tag{1}$$

where $|I|$ is the number of the images in the dataset.

Step 2: Geometric verification. Although the top ranked retrieved images from step one have similar visual appearance in terms of their bag-of-words representation, they do not necessarily share the same geometric similarity as well:

$$\text{when } d\left(I_Q, I^j_Q\right) \text{ is small} \nRightarrow g\left(I_Q, I^j_Q\right) > \gamma. \tag{2}$$

Images that are ranked highly according to bag-of-words but they exhibit a poor geometric similarity are considered geometrically inconsistent. We simply filter out these geometrically inconsistent retrieved images. In order to minimize the possibility of false geometric transformations, we impose harsh geometric constraints, that is we set the threshold γ high. For computational reasons, we limit the number of geometric checks to the top M retrieved images. At the end of this step, we have per-query the assumed positive j images where $1 \leqslant j \leqslant M$ and their geometric transformations $H\left(I_Q, I^j_Q\right)$ with respect to the query image.

Step 3: Visual synonym candidates. Based on the estimated geometric transformations from step 2, we seek for word pairs of *Type* 3. We do so by back-projecting the geometry transformation $H\left(I_Q, I^j_Q\right)$ onto I_Q and I^j_Q. Then, we search for word pairs $p_{r,t} = \{w^r, w^t\}$ that belong to pair of features under the condition of *Type* 3, that is

$$p_{r,t} = (w^r_{k,I_Q}, w^t_{l,I^j_Q}) : |\mathcal{X}_{k,I_Q} - H(I_Q, I^j_Q) \cdot \mathcal{X}_{l,I^j_Q}| < \epsilon \tag{3}$$

where k and l iterate over all features in images I_Q and I^j_Q respectively and ϵ is a user defined variable. At the end of this step, we have a list of pairs of visual synonym candidates $\mathcal{P} = \{p_{r,t}\}$.

Step 4: Removal of rare pairs. Although we employ harsh geometric constraints, false positive geometry estimation is still possible to happen. In that case the word pairs harvested are incorrect and they should be filtered out.

Assuming that false positive geometry estimation is not repetitive over specific images, the visual word pairs that subsequently arise from these pairs of images do not appear frequently. Hence by applying a frequency threshold we are able to exclude these presumably false visual word pairs. Moreover this frequency threshold conveniently reduces the number of visual synonym word pairs to a manageable size. The occurrence frequency of all pairs of visual synonym candidates is thresholded at f to drop word pairs that occur too rarely. The final list of synonyms is composed of the pairs

$$\mathcal{S} \subset \mathcal{P} : f_{p_{r,t}} > f \tag{4}$$

We summarize the algorithm in Fig. 3.

3.4. Landmark image retrieval using visual synonyms

Visual synonyms are pairs of words, ideally originating from the near identical physical parts of reality. It is therefore expected that visual synonym words will appear in images of the same landmarks. For example two visual synonym words that depict the tip of Eiffel tower will appear in Eiffel tower images with high frequency. If we retain those visual synonym words then, we expect that they will supplement each other in these landmark scenes in which they frequently appear.

Having harvested the visual synonyms, we are equipped with a pool of words $\mathcal{S}$ that could potentially participate in the construction of a new reduced vocabulary. In the current work the new reduced vocabulary of size $|R|$ is constructed by selecting the words participating in the most frequent visual synonym pairs. Because visual synonyms are pairs of words, we use the top $|R|/2$ visual synonyms, that is

$$R = p_{r,t} : f_{p_{r,t}} > f^{|R|/2}_{p_{r,t}}, \tag{5}$$

where $f^{|R|/2}_{p_{r,t}}$ is the frequency of the $|R|/2$ most frequent visual synonym pair. Practically, words appear in more than one visual synonym pair, therefore the final vocabulary size is typically smaller than $|R|/2$.

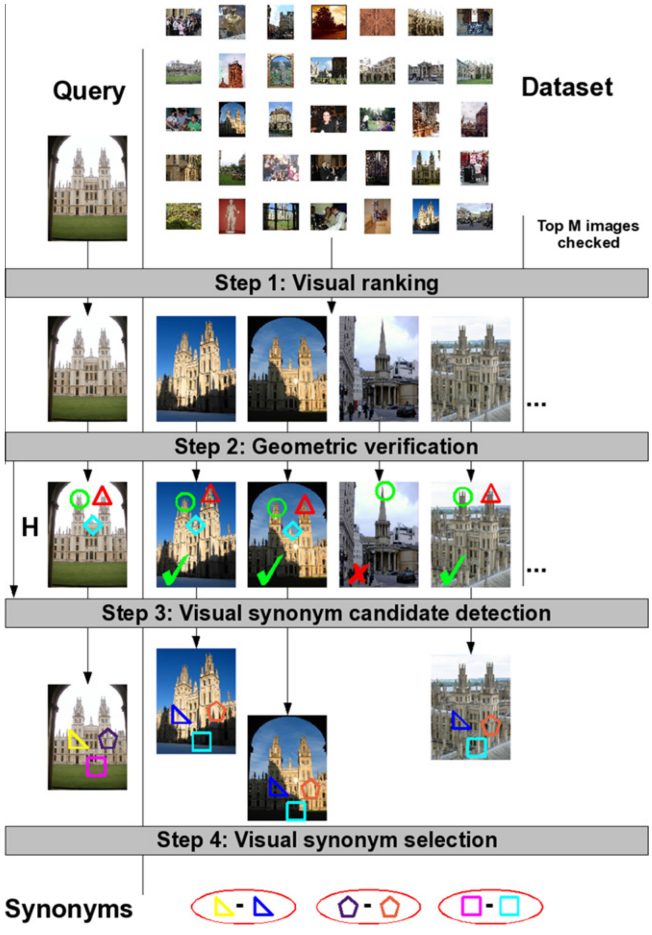

Query
Dataset
Top M images checked
Step 1: Visual ranking
...
Step 2: Geometric verification
H
...
Step 3: Visual synonym candidate detection
Step 4: Visual synonym selection
Synonyms

Fig. 3. The 4-step algorithm for finding visual synonyms. First, we rank images according to their bag-of-words similarity with the query image. Second, we select from the ranked images the most likely positives by using the number of geometric inliers (features with same color and shape, e.g. ○–○). Then, by using the homography matrix H we back-project those words, that are assigned to different clusters but reside in the same physical image location (features with the same shape but different color, e.g. □–□). These are the visual synonyms candidates. Finally, after repeating the procedure for all the query images, we use a threshold to maintain only frequent visual synonym candidates.

3.5. Landmark image retrieval with synonym augmentation

We employ the visual augmentation model for updating the image histogram representation according to the reduced vocabulary of the visual synonyms words. We closely follow the approach proposed in [25,24]. However, we simplify the model and make no use of the geometric verification step. We consider only the first retrieved image as a positive. We then average the histograms of the query image and the top retrieved image. The averaged histogram is our new query. For the bag-of-words retrieval, that is given a query image, we search for the closest image based on a predefined distance measure. Naturally, the top retrieved image will again be retrieved in the top rank.

3.6. Implementation

3.6.1. Bag-of-words representation

We experiment with two different types of descriptors. First, we describe Hessian-Affine detected keypoints with SIFT [1,5]. Second, we use the SURF descriptor [2] and the detector with which it was proposed. For both cases we use 200K vocabularies, trained on an independent 30K dataset downloaded from Flickr. Because of the large size of the vocabulary, we use approximate nearest neighbor techniques for the word assignment stage. For this purpose, we rely on the FLANN library [30]. Finally, we use histogram intersection as visual similarity measure to rank the images in Step 1.

3.6.2. Geometry

We perform the geometric verification on the top $M = 40$ images. Although in the literature a value of $\gamma = 25$ inliers is considered to be an acceptable threshold for accepting a pair of images as geometrically consistent [31], we require particulary high precision to avoid false positives. The higher we set threshold γ the smaller the amount of geometrically verified images is. If we set the threshold too high, for example to $\gamma = 100$ inliers, only pairs of near duplicate images would be considered as geometrically consistent. We set the minimum required number of inliers to $\gamma = 50$, a value which was empirically found to result in high precision (data not shown). We estimate the homography matrix using the fast spatial matching algorithm introduced in [14]. For homography matrix estimation the symmetric transfer error function is used as cost function [26]. The maximum distance error then is taken $\delta = 0.001$ and the approximation error $\epsilon = \delta/10$.

4. Experimental setup

4.1. Data set

We report our experiments on the *Oxford5k* data set [14], which contains 5062 large Flickr images from 11 landmark scenes in Oxford. The images are labeled either as “good”, “ok” or “junk”, depending on how clear is the view of the scene. When a picture depicts clearly the scene, it is labeled as “good”, whereas when more than 25% of the scene is visible inside the picture, then the image is labeled as “ok”. Images in which less than 25% of the object is visible are labeled as “junk”. The number of pictures labeled as “good” or “ok” differs from scene to scene, ranging from as few as 6 images for “Pitt Rivers” to ∼200 images for “Radcliffe Camera”. To simulate a real word scenario, *Oxford5k* contains more than 4000 additional images that depict none of the landmarks. The landmark images in *Oxford5k* are known to contain many occlusions as well as large changes in scale, viewpoint, and lighting conditions. For all these reasons, *Oxford5k* is a challenging dataset. The algorithm is fully unsupervised, therefore the data are not split into training and test set.

4.2. Visual synonyms extraction

Visual synonym extraction is an unsupervised technique, since there is no need for prior knowledge of the true similarity between landmark images. Instead, geometry and harsh geometrical constraints provide us with this type of information. Despite the harsh geometric constraints, there are still pairs of non-landmark images, which are successfully matched.

4.3. Experiments

4.3.1. Visual synonym validity

A direct quantitative evaluation of visual synonyms is hard, since no ground truth of individual patch semantics, such as “window” or “doric column”, is available. Instead, we use the landmark ground truth label for the entire image in which a visual synonym word is found. Given a visual synonym word pair, we count how many times these visual synonym words appear in the same landmark images. We repeat using the same visual synonym words and random words. We then compare against the landmark cooccurrence of the visual synonyms words. This is an indirect measurement, since the fact that two visual words appear to the same landmark does not necessarily imply that they also depict the same visual reality. However, we ignore this possibility and hypothesize that a pair of visual synonym words should appear more frequently in common landmarks, than random pairs of words.

4.3.2. How close are visual synonyms in descriptor space?

An important property that would reveal the potential of visual synonym words is the distribution of their distances in the feature space. Given an $L2$-normalized feature space forming a sphere with unit length radius, all features lie on the surface of the sphere and visual words form Voronoi cells. The distance between every pair of words is proportional to the angle of the corresponding visual word vectors. When visual synonym words are close, compared to random words, their vectors angle is smaller than the angle between the random word vectors. We test whether visual synonyms are most often pairs of neighboring cells or distant cells. To examine we calculate the distribution of distances in experiment 1.

This experiment operates in feature space, which in our case is the 128-D SIFT, or 64-D SURF space. To answer this question, we calculate the distances $d_{r,t} = d(w^r, w^t)$, $d_{r,j} = d(w^r, w^j)$ and $d_{t,j} = d(w^t, w^j)$ for $w^j j \neq r,t$. We use cosine similarity distance, that is $c(w^r, w^t) = \frac{\sum_i x_i^r \cdot x_i^t}{|x^r| \cdot |x^t|}$, where x_i^r is the ith coordinate of the feature vector of w^r. Next, given a word we rank the distances from the rest of the words and mark the distance from its visual synonym word. Closer visual synonym words would imply lower ranks.

4.3.3. Vocabulary reduction using visual synonyms

Next, we want to study whether visual synonyms can successfully be used for vocabulary reduction, despite the instability of the feature space. If visual synonyms are repeatable enough, so that they retain the vocabulary's performance levels, this would also imply that transforming them to visual metonyms is a reasonable option. We select a subset of the visual synonyms extracted and use them as a new visual vocabulary. In this experiment, the reduced vocabularies range from 30 K words to 1 K words. We compare against reduced vocabularies based on visual metonyms, similar to the reduction method proposed by Turcot and Lowe [25], and the full 200 K vocabulary baseline. We also compare against a reduced vocabulary derived from the combination of visual synonyms and metonyms.

4.3.4. Landmark image retrieval using visual synonyms and visual augmentation

Finally, we study whether the visual synonym reduced vocabularies are orthogonal to other, state-of-the-art retrieval techniques. To this end we make use of visual augmentation [24]. In the current setup, we simplify visual augmentation by considering the first retrieved image I_Q^1 to belong to the same landmark as the query image. We then use I_Q^1 to update the query image histogram.

4.4. Evaluation protocol

For the evaluation of our retrieval experiments, we follow the protocol suggested in [14]. Five images from each landmark are used as query images and the average precision scores are averaged over each landmark scene for the final result.

Our evaluation criterion for this retrieval experiment is the average precision score:

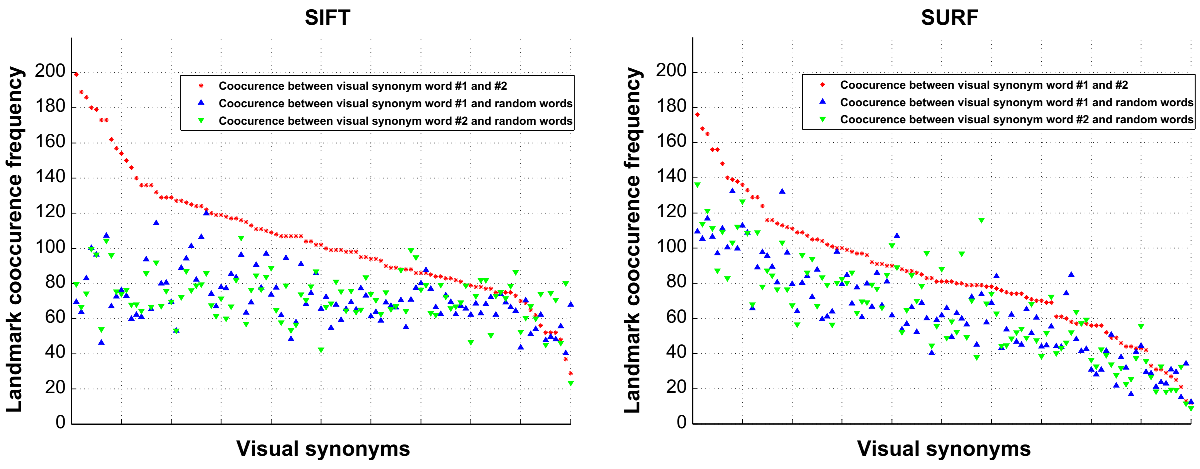
$$AP = \frac{\sum_{j=1}^{N} P(j) \cdot R(j)}{\sum_{j=1}^{N} R(j)}$$

where j is the rank, N is the total number of documents, $P(j)$ is the precision at the jth rank and $R(j)$ is a binary function, which is true if the jth retrieved document is a true positive result.

5. Results

5.1. Experiment 1: Visual synonym validity

We plot the results for experiment 1 in Fig. 4a. In the majority of the cases visual synonyms words cooccur in similar landmarks images more often than random words for both (a) SIFT and (b) SURF. More specifically, visual synonyms words not only cooccur in the same landmarks, but also have similar visual appearance. In contrast, random words at the tail of both Fig. 4a and b cooccur in the same landmarks with visual synonym words but depict different visual details. Nonetheless, the tail of the distribution highlights the pairs of visual synonym words, were the number of potential matches between individual patches of the respective words is smaller. This number of potential matches corresponds to the confidence of the visual synonym pair, i.e. the more potential matches, the higher the confidence.



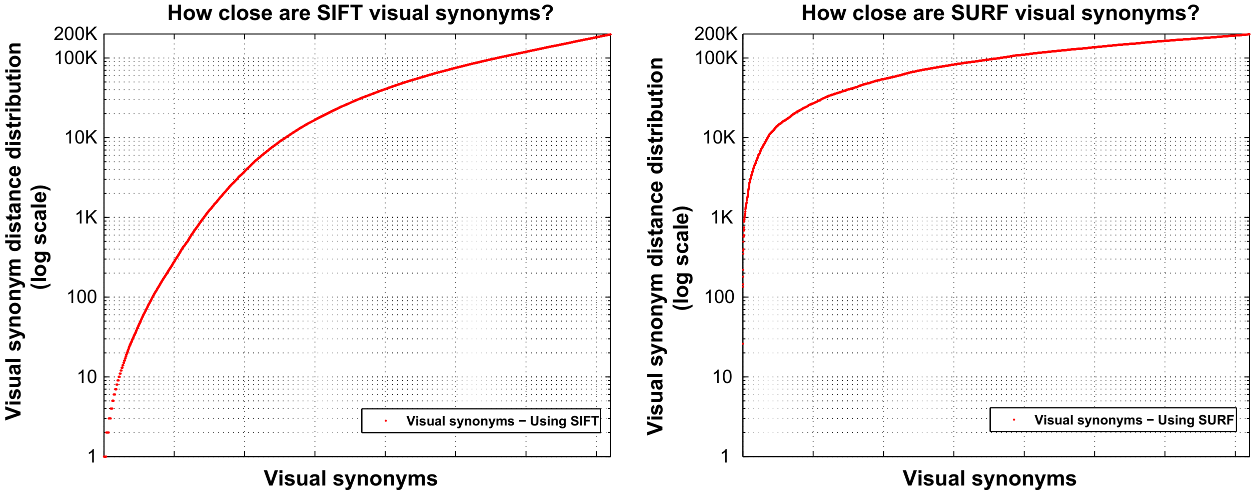
(a) (b)

Fig. 4. Results from experiment 1. Cooccurrence of landmarks between visual synonym words and random words using (a) SIFT and (b) SURF.

5.2. Experiment 2: How close?

We show the results of experiment 2 in Fig. 5. Each dot corresponds to the distance between two visual synonym words, w_{s_1}, w_{s_2}. The y-value relates to the comparison of the distance $\langle w_{s_1}, w_{s_2}\rangle$ and $\langle w_{s_1}, w_j\rangle$ for all j except for s_2. The smaller the distance between the synonym words compared to random words, the lower the dot is and therefore the closer the synonym words are, as compared to all other words. Hence, the figure provides us with the distribution of distance rankings between visual synonym words and all other words. Naturally, there are many visual synonym pairs that lie close in feature space (visual synonyms in the top 100 closest words). However, the spectrum of distance rankings is broad. While some synonyms are relatively close neighbors indeed, lying for example in the range 0–100 in Fig. 5, the majority of word pairs tends to be distant from one another. This shows that visual synonym words are scattered throughout descriptor space, regardless their common origins from the same physical elements in the 3D scenes. If we added in the current plot also visual metonyms, the plot would acquire a sigmoid shape, since the metonyms would by definition be the closest words in feature space.

For visual synonyms extracted with the SURF descriptor we observe a much steeper curve, see Fig. 5b. We attribute this steeper



(a) Visual synonym cluster distances using SIFT

(b) Visual synonym cluster distances using SURF

Fig. 5. Results from experiment 2. Each red dot stands for a visual synonym word pair. The lower the dot is, the closer the corresponding visual words are. We sort the dots from lower to higher. Hence, the lines show the distribution of closeness between visual synonyms. For both SIFT and SURF features, the visual synonyms extracted are not that close. Although there are pairs of words that lie close enough, for example in the range 0–100, there are also many word pairs that lie far enough in the feature space. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

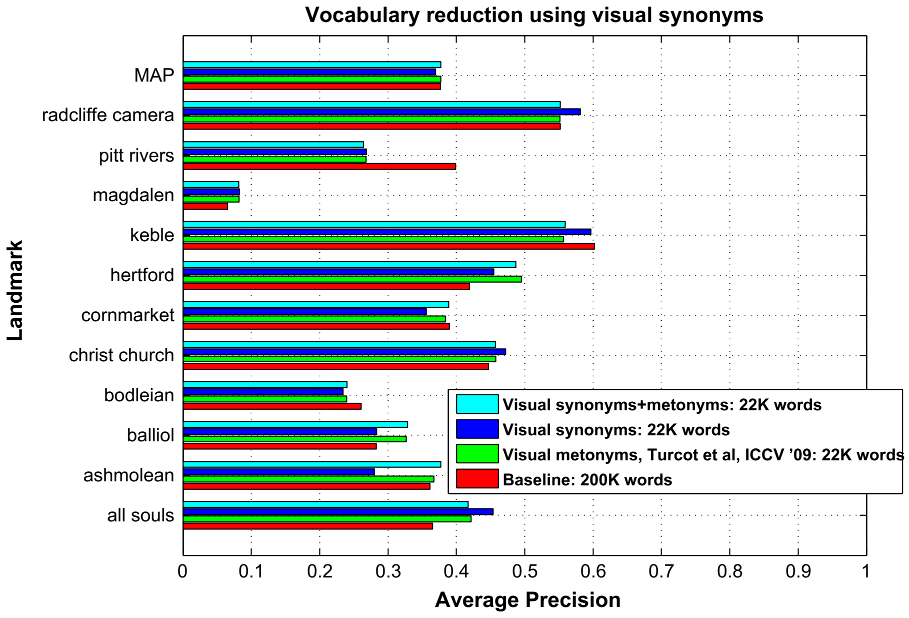


Fig. 6. Results from experiment 3. Visual synonyms construct reduced vocabularies that perform the same with the state-of-the-art. Using the 22 K visual synonyms vocabulary, that is 89% compression rate, we perform the same as the 200 K baseline. Although visual synonyms are associated with the instability of the feature space and the word assignment process, visual synonym reduced vocabularies perform on par with the more stable visual metonyms [25] and the baseline. Combining visual synonyms and metonyms leads to a small performance increase.

curve to two reasons. First, fewer visual synonym statistics are available. Contrary to SIFT, where we obtain 5409 features per image on average, using SURF we have 1900 features on average. As a result, there are fewer visual word matches between retrieved images. Hence, the estimated homographies are not as precise. Since the proposed algorithm largely depends on having accurate homography mappings from one image to the other, visual synonym statistics from SURF features are not adequate. Thus visual synonym extraction is not as reliable. For the same reason the number of visual synonyms extracted using SURF is much smaller: 7263 visual synonym word pairs when using SURF whereas 72,337 visual synonym word pairs when using SIFT. Second, the SURF detector does not find elliptical shaped features, like the Hessian-Affine feature detector that we use for SIFT. Therefore patches are not affine normalized before extracting the SURF features, which is a disadvantage for landmarks.

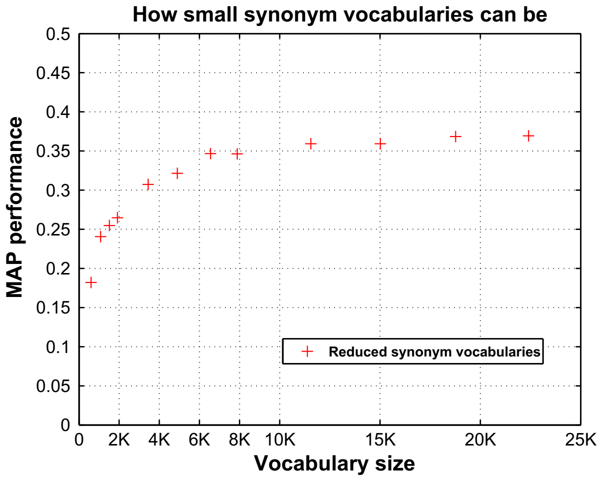


Fig. 7. How small synonym vocabularies can be? Compact vocabularies using visual synonyms can be reduced to 2–4 K words, a 99% compression rate, with no severe degradation on MAP (5% decrease).

5.3. Experiment 3: Vocabulary reduction using visual synonyms

We show the results of experiment 3 in Fig. 6. The size of the reduced vocabularies ranges from 1 K, a 99.5% reduction ratio to 22 K, a 89% reduction. The 22 K visual synonym vocabulary performs the same as the full vocabulary of 200 K words. Hence, using a 89% smaller vocabulary, we are able to achieve similar performance. When constructing a 22 K vocabulary based on visual metonyns, essentially following the approach introduced by Turcot and Lowe [25], the performance remains on similar levels. The same performance is obtained when visual synonyms and metonyms are both used for the construction of the reduced vocabulary. The number of the words found to participate both in the visual synonyms and visual metonyms vocabulary is 11 K. Naturally, visual metonyms are more consistent and therefore more robust. Although visual synonyms are associated with the instability of the feature space and the word assignment process, the fact that the respective vocabularies perform on par indicates that visual synonyms carry useful information.

In Fig. 7 we plot the performance of many reduced vocabularies, ranging from 1 K to 22 K words. The performance of the vocabularies remains close to baseline levels for as much as 98–99% smaller vocabularies (3–4 K), after which there is a noticeable performance degradation (more than 5% in MAP). Consistent performance with 98% smaller vocabularies shows that the redundancy initially introduced by large vocabularies is minimized.

In the above analysis, there are several parameters that are manually set, such as M and ϵ. Varying M or ϵ does not lead to any significant changes in performance (data not shown). We therefore conclude that visual synonym extraction is robust to small parameter deviations.

5.4. Experiment 4: Landmark image retrieval using visual synonyms and visual augmentation

We show the results of experiment 4 in Fig. 8. We used two visual synonym vocabularies, the 4 K and the 22 K. We consider as the baseline the standard bag-of-words model using a visual vocabulary of size 200 K. Using the 4 K reduced vocabulary

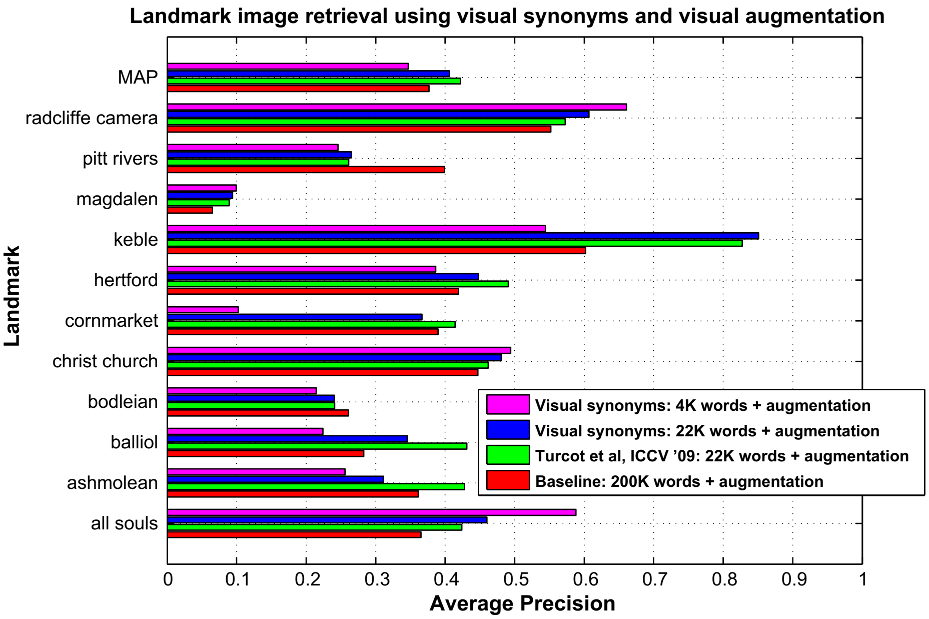


Fig. 8. Results from experiment 4. Using visual augmentation of the visual synonym reduced vocabulary increases performance to 0.406 in MAP. Using the 22 K vocabulary with visual augmentation [25], the performance is the same. Using the 4 K vocabulary with visual augmentation [25], results in a performance of 0.347, a loss of only 3% compared to the baseline.

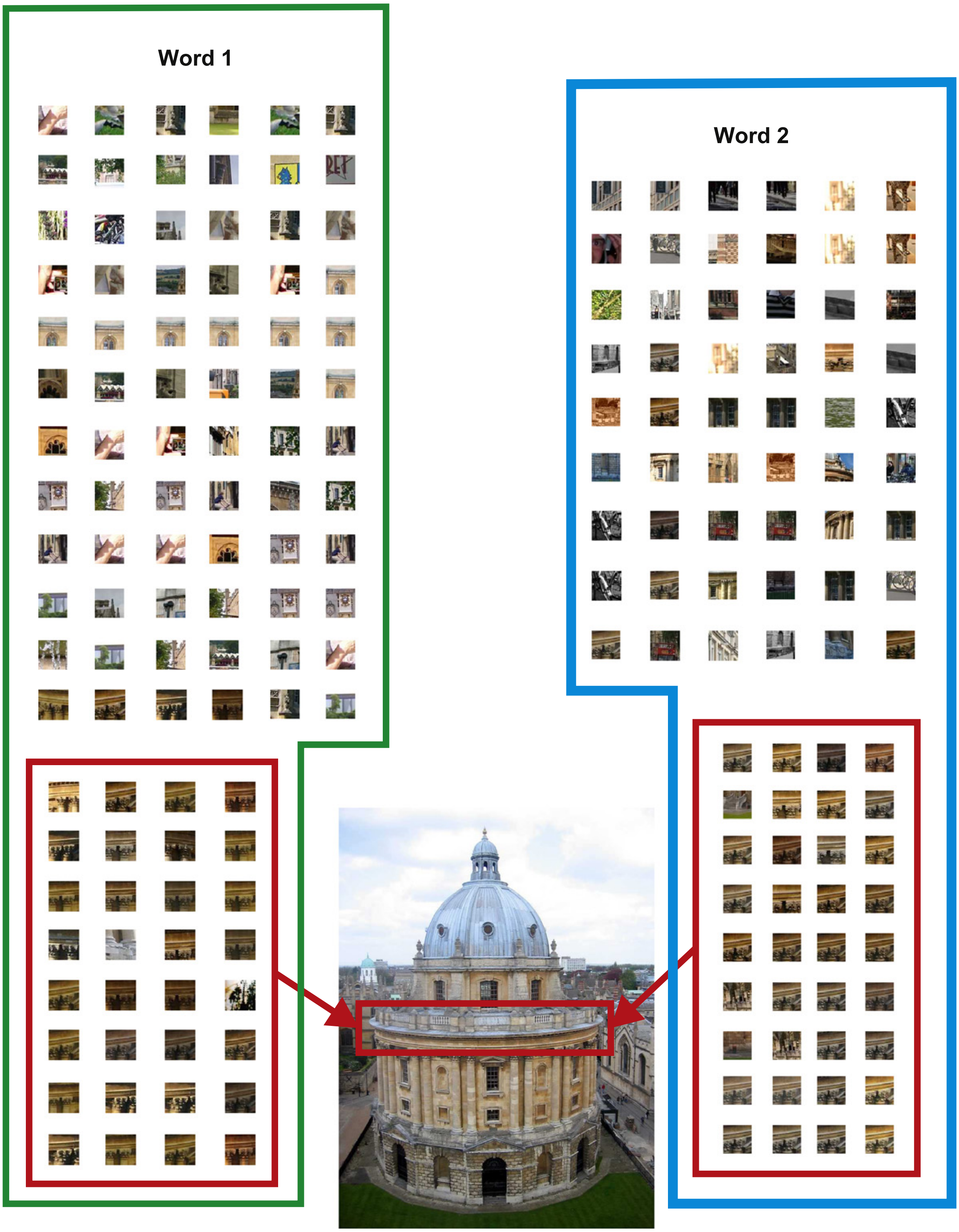


Fig. 9. Example of visual synonym words. The green and blue polygons enclose patches from two different visual words found to be visual synonyms. The two red rectangles focus on patches of each visual word separately. The resemblance between the patches inside the red rectangles proves why the two words were labeled as synonyms. At the bottom of the figure we show a picture of the "Radcliffe Camera" scene, indicating the location where the most visual word patches are extracted. Patch sizes scaled to fit the page. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

together with visual augmentation we perform almost the same as the baseline of the 200 K words (0.347 MAP for 4 K vocabulary with visual augmentation vs 0.377 for full 200 K vocabulary). When using the top performer 22 K vocabulary, the performance increases by 4%, going to 0.406 in MAP.

In Fig. 11 we highlight retrieval results using the reduced visual synonym vocabularies. The query image is a picture of the "All souls" college. When we follow the standard baseline approach, first row, we score a poor 0.154 in average precision. We attribute this low score to the noticeable visual deformations that the query image undergoes, that is extreme illumination variation within the same picture. Hence, it is no surprise that the baseline retrieves two wrong pictures in the top ranked results. In the second row we highlight the landmark image

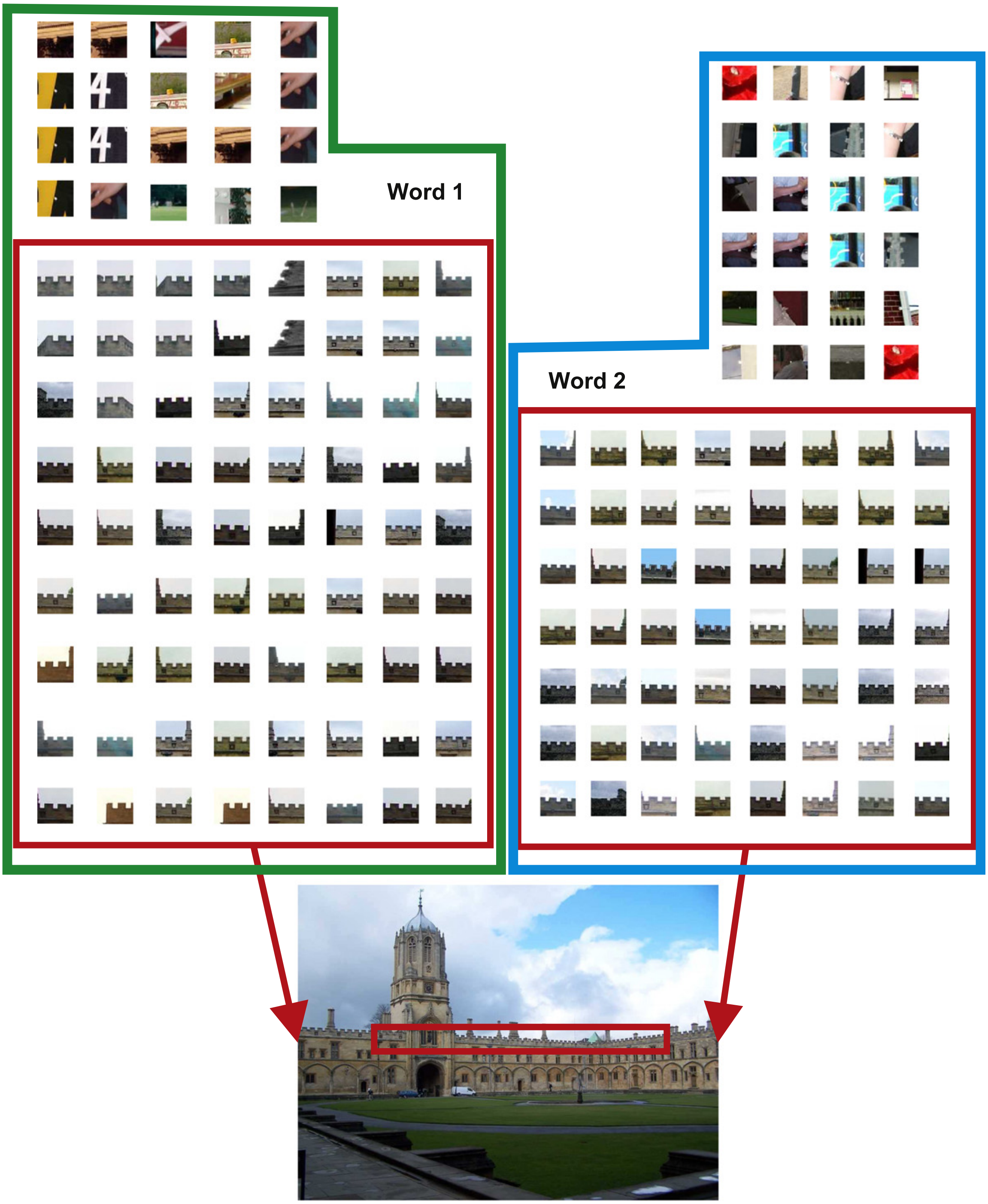
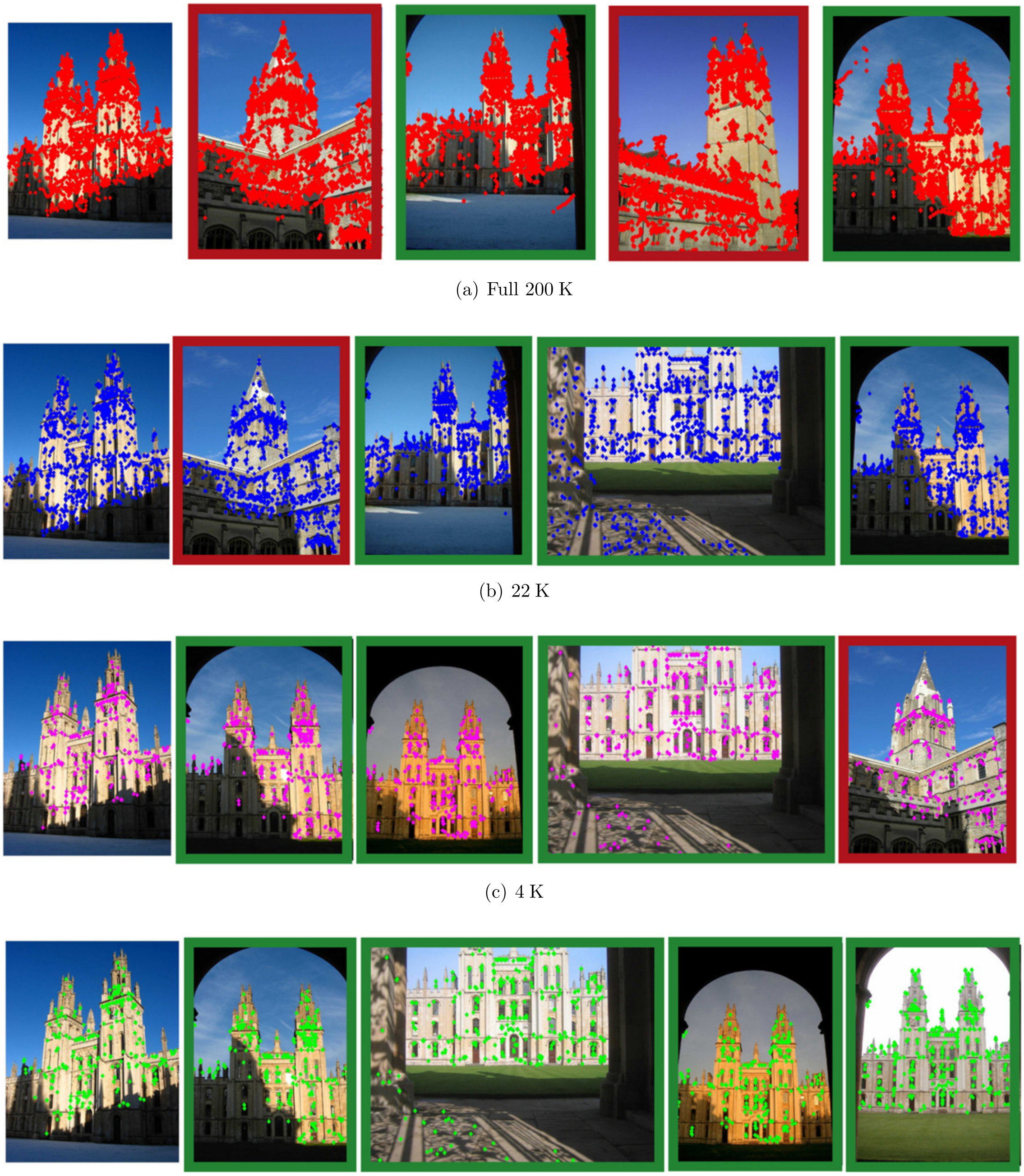


Fig. 10. Another example of visual synonym words. Again, The green and blue polygons enclose patches from two different visual words found to be visual synonyms. The two red rectangles focus on patches of each visual word separately. The resemblance between the patches inside the red rectangles proves why the two words were labeled as synonyms. At the bottom of the figure we have a picture of the "Christ church" scene, indicating the location where the most visual word patches are extracted. In this example becomes clear why these words are found to be visual synonyms, since they both share a characteristic castle loophole pattern. Patch sizes scaled to fit the page. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

retrieval results using visual synonyms with a 22 K vocabulary. As can be observed, the false positive ranked at position one when using the baseline, is maintained, however the false positive in rank 3 is substituted from a true positive. It is the reason why the average precision jumps from 0.154 to 0.224. For retrieval using the 4 K visual synonym vocabulary the results further improve to an average precision of 0.236. It is worthwhile mentioning that for the reduced vocabularies using visual synonyms, the detected visual words (blue and magenta dots respectively) are located on the visual elements, that are characteristic of the landmark, that is the towers and the gothic style textured windows. The last row shows the retrieved results when using the 4 K vocabulary together with visual augmentation. Using visual augmentation further pushes the performance to 0.399. This is reflected to the retrieval, where the four top ranked images are all true positives.



(a) Full 200 K

(b) 22 K

(c) 4 K

(d) 4 K with visual augmentation

Fig. 11. Retrieval results for an "All souls" query image, using three different vocabularies. The full 200 K vocabulary, our baseline, corresponds to the first row. In the second row we have the 22 K visual synonym vocabulary. In the third row we have the 4 K visual synonym vocabulary. Finally, in the last row we again have the 4 K visual synonym vocabulary combined with visual augmentation. The visual words extracted are depicted as colored dots. For the reduced vocabularies, the visual words are mainly spotted in geometrically reasonable positions.

5.5. Qualitative results

Examples of visual synonym words are illustrated in Figs. 9 and 10. In both figures the red rectangles reveal why those two visual words have been selected as visual synonyms. These patches come from the "Radcliffe Camera" and "Christ Church" landmark respectively. Under the bag-of-words model each visual word w_j covers a sensitive subspace $\mathcal{F}_{w_j}$ in descriptor space $\mathcal{F}$. Although parts of $\mathcal{F}_{w_j}$ include coherent patches, there are clearly many more patches that reside in the same part of the feature space and are visually different. As a result, we have visual synonyms, that is features arising from the same 3D elements of the physical world, yet assigned to different and probably distant visual words.

6. Conclusions

In this paper we have introduced the notion of visual synonyms, which are independent visual words that nonetheless cover similar appearance. To find synonyms we use conspicuous elements of geometry on landmarks. Different views of the same conspicuous element are usually covered by different parts of the feature space; they are the synonyms to be. We detect pairs of synonyms by mapping the conspicuous elements onto one another. We evaluate the validity of visual synonyms. They appear in consistent landmark locations. Using SIFT descriptors seems to yield visual synonyms of better quality than using SURF, because the latter generates a smaller number of features on average. We tested visual synonyms with respect to their closeness in descriptor space. They appear to be not just simply close neighbors, even when they have a similar appearance. In fact they can be very far in feature space.

Visual synonyms can be used for vocabulary reduction, obtaining 98–99% smaller vocabularies with only 5% performance degradation in MAP. Furthermore, combination of visual augmentation together with 98–99% smaller visual synonym vocabularies boosts performance to baseline levels. The reduction achieved demonstrates that visual synonyms carry useful information. Although arising from the inconsistency of the feature space and the word assignment process, visual synonyms capture the essence of the landmark scenes of interest.

All in all, visual synonyms provide a look into the multidimensional feature space, allow us to study the nature of the visual words and their intrinsic incoherence, maintain landmark image retrieval performance, and in the end reduce the visual vocabulary size.

Acknowledgment

This research is supported by the IM-Pact BeeldCanon project.

References

[1] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.

[2] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Understand. 110 (3) (2008) 346–359.

[3] A.F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and trecvid, in: Proceedings of the 8th ACM international workshop on Multimedia information retrieval, 2006, pp. 321–330.

[4] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[5] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: a survey, Found. Trends Comput. Graph. Vis. 3 (2008) 177–280.

[6] K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.

[7] K.E.A. van de Sande, T. Gevers, C.G.M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.

[8] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, Int. J. Comput. Vis. 60 (1) (2004) 63–86.

[9] J. Matas, O. Chum, U. Martin, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, in: British Machine Vision Conference, vol. 1, 2002, pp. 384–393.

[10] T. Tuytelaars, L. Van Gool, Wide baseline stereo matching based on local, affinely invariant regions, in: British Machine Vision Conference, 2000, pp. 412–425.

[11] T. Tuytelaars, L. Van Gool, Matching widely separated views based on affine invariant regions, Int. J. Comput. Vis. 59 (1) (2004) 61–85.

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (2005) 43–72.

[13] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: International Conference on Computer Vision, vol. 2, 2003, pp. 1470–1477.

[14] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[15] J.R.R. Uijlings, A.W.M. Smeulders, R.J.H. Scha, What is the spatial extent of an object? in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 770–777.

[16] S. Zhang, Q. Tian, G. Hua, Q. Huang, S. Li, Descriptive visual words and visual phrases for image applications, in: Proceedings of the ACM International Conference on Multimedia, 2009, pp. 75–84.

[17] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Neural Information Processing Systems, 2004.

[18] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2161–2168.

[19] Z. Wu, Q. Ke, M. Isard, J. Sun, Bundling features for large scale partial-duplicate web image search, 2009, pp. 25–32.

[20] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J.-M. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.

[21] E. Gavves, C.G.M. Snoek, Landmark image retrieval using visual synonyms, in: Proceedings of the ACM International Conference on Multimedia, 2010, pp. 1123–1126.

[22] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[23] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: British Machine Vision Conference, 2011.

[24] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: Automatic query expansion with a generative feature model for object retrieval, in: International Conference on Computer Vision, 2007.

[25] P. Turcot, D.G. Lowe, Better matching with fewer features: The selection of useful features in large database recognition problems, in: International Conference on Computer Vision, 2009.

[26] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.

[27] H. Jegou, M. Douze, C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in: European Conference on Computer Vision, 2008, pp. 304–317.

[28] G. Schindler, M. Brown, R. Szeliski, City-scale location recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–7.

[29] Y.S. Avrithis, Y. Kalantidis, G. Tolias, E. Spyrou, Retrieving landmark and non-landmark images from community photo collections, in: Proceedings of the ACM International Conference on Multimedia, 2010, pp. 153–162.

[30] M. Muja, D.G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration, in: International Conference on Computer Vision Theory and Applications, 2009, pp. 331–340.

[31] O. Chum, J. Matas, J. Kittler, Locally optimized RANSAC, Pattern Recogn. (2003) 236–243.