ABSTRACT
In this paper, we consider the incoherence problem of the visual words in bag-of-words vocabularies. Different from existing work, which performs assignment of words based solely on closeness in descriptor space, we focus on identifying pairs of independent, distant words – the visual synonyms – that are still likely to host image patches with similar appearance. To study this problem, we focus on landmark images, where we can examine whether image geometry is an appropriate vehicle for detecting visual synonyms. We propose an algorithm for the extraction of visual synonyms in landmark images. To show the merit of visual synonyms, we perform two experiments. We examine closeness of synonyms in descriptor space and we show a first application of visual synonyms in a landmark image retrieval setting. Using visual synonyms, we perform on par with the state-of-the-art, but with six times less visual words.

Categories and Subject Descriptors
I.2.10 [Vision and scene understanding]: Vision

General Terms
Algorithms, Measurement, Experimentation

Keywords
Visual words, synonyms, geometry, landmark retrieval

1. INTRODUCTION
The bag-of-words model is a method inspired by text retrieval, which has been applied in a variety of visual retrieval and categorization contexts [5, 6, 8, 11]. The basic idea behind the model is to view an image as a document, by treating local image descriptors as orderless words. We obtain words by clustering [4] the descriptor space, and simply assume that different clusters correspond to different visual words. In contrast to text retrieval, however, no clearly defined words exist in the visual domain. Consequently, the most challenging part of the bag-of-words model is to acquire a meaningful vocabulary of distinctive visual words.

When we consider the typical visual words resulting from clustering in Figure 1, we observe that similar patches are not necessarily assigned to the same cluster. What is more, some clusters appear clearly incoherent. Starting from these observations, we explore in this paper whether the visual word representation in the bag of visual words model can be improved. To study the problem, we focus on landmark images [1, 5] that are characterized by their constant geometry. We make a first attempt to connect different visual words, resulting from clustering, based on their geometric appearance. We call these connected words visual synonyms.

The bag-of-words method is the state-of-the-art approach in landmark image retrieval [5]. An efficient and cheap extension is “visual augmentation” [1, 7]. Visual augmentation updates the query image histogram based on the query’s closest neighbors histograms. 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. In this paper we will use an approach similar to visual augmentation to examine the effect of visual synonyms in landmark image retrieval.

2. VISUAL SYNONYMS
We define visual synonym words as “independent visual words, which host descriptors representing image patches with similar visual appearance”. Nonetheless, these words contain descriptors that correspond to image patches originating from the very same physical element.

To obtain visual synonyms we must find different visual words that are likely to host visually similar patches. We

![Figure 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.](image-url)
cannot rely on image appearance only, since it is the cause of the problem. Therefore, we need an independent information source to supply us with additional visual knowledge. For landmark images, containing pictures of the same physical locations, the use of geometry makes most sense as the scene remains largely unchanged [9].

2.1 Preliminaries

We first introduce some notation for the ease of explanation. Following the query-by-example paradigm, we refer to a query image of dataset \( I \) as \( I_Q \) and to the rest of the ranked images ranked as \( I_Q^j \), where \( j \) denotes the rank of the retrieved image. We define image feature \( \xi \) as the local descriptor extracted on an interest keypoint with scale and location \( X \), mapped to a visual word \( w' \) of the vocabulary. Consequently, the \( i \)-th feature of image \( I_1 \) is denoted as \( \xi_{1,i} = \{ w_{1,i}, X_{1,i} \} \). Finally, two images \( I_Q \) and \( I_Q^j \) are geometrically connected with a homography matrix \( H(I_Q, I_Q^j) \).

2.2 Connecting visual words with geometry

Two images are connected with a matrix \( H \), which is estimated using RANSAC [2]. Since RANSAC needs one to one point correspondences and given the visual features and their unique spatial locations in the two images, four possible feature pair relations exist, see also Figure 2.

**Type 1**: features \( \xi \) that are mapped to the same visual words \( w \) and lie in consistent physical locations, that is
\[
\xi_{1,i}, \xi_{2,j} : w_{1,i} = w_{2,j}, X_{1,i} \approx H(I_1, I_2) \cdot X_{2,j}.
\]

**Type 2**: features \( \xi \) that are mapped to the same visual words \( w \) and lie in different physical locations, that is
\[
\xi_{1,i}, \xi_{2,j} : w_{1,i} = w_{2,j}, X_{1,i} \neq H(I_1, I_2) \cdot X_{2,j}.
\]

**Type 3**: features \( \xi \) that are mapped to different visual words \( w \) and lie in consistent physical locations, that is
\[
\xi_{1,i}, \xi_{2,j} : w_{1,i} \neq w_{2,j}, X_{1,i} \approx H(I_1, I_2) \cdot X_{2,j}.
\]

**Type 4**: features \( \xi \) that are mapped to different visual words \( w \) and lie in different physical locations, that is
\[
\xi_{1,i}, \xi_{2,j} : w_{1,i} \neq w_{2,j}, X_{1,i} \neq H(I_1, I_2) \cdot X_{2,j}.
\]

Feature pairs of Type 1 and Type 2 are widely used in the literature as input to RANSAC [7]. Naturally, feature pairs of Type 4 make less sense, 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 appearance, a property which state-of-the-art landmark image retrieval and classification methods [1,5,8] fail to capture. Therefore, we focus on the pairs of visual words of the feature pairs Type 3 to study the visual word incoherence.

2.3 Visual synonyms extraction

Our visual synonym extraction algorithm is a three-step procedure. We use two different distance measures: a visual similarity distance measure \( d(\cdot) \) and a geometric similarity measure \( g(\cdot) \). For visual similarity distance measure, either cosine similarity, standard euclidean distance or histogram intersection are usually chosen. As a form of geometric similarity, typically, the number of inliers between two images returned from RANSAC is used [1]. We introduce a geometric threshold \( \gamma \), which refers to the minimum number of inliers returned from RANSAC, which we use to judge whether two images are geometrically related.

**Step 1: Visual ranking** We rank all images in a data set according to their visual similarity with respect to a query image \( I_Q \), using the standard bag-of-words model for modelling visual appearance. After this step, we obtain an ordered list \( \{ I_Q, I_Q^1, \ldots, I_Q^{|I|} \} \), such that:
\[
d(I_Q, I_Q^j) < d(I_Q, I_Q^{j+1}), \quad j = 1, \ldots, |I| - 1,
\]
where \(|I|\) is the number of the images in the dataset.

**Step 2a: 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 have the same small geometric distance as well:
\[
d(I_Q, I_Q^j) \approx 0 \neq g(I_Q, I_Q^j) > \gamma.
\]

We use image geometry to filter out the inconsistent retrieval results. After the geometric verification, we consider all the retrieved images relevant with respect to the query image and suitable for visual synonym extraction. Therefore, we impose harsh geometric constraints to minimize the possibility of false geometric transformations. 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 images and their geometric transformations \( H \) with respect to the query image.

**Step 2b: Visual synonym candidate detection** For each query image, we hold a list of assumed positive images and their geometric transformation to \( I_Q \). Based on these estimated geometric transformations, we seek for word pairs of Type 3. We do so by back-projecting the geometry transformation \( H \) between \( I_Q \) and \( I_Q^j \) and searching for pairs of words \( p_{r,t} \) belonging to feature pairs of Type 3, that is
\[
p_{r,t} = \{ w_{1,Q,k}, w_{1,t} \} : X_{1,Q,k} \approx H(I_Q, I_Q^j) \cdot X_{1,t}.
\]

where \( k, l \) iterate over all features in \( I_Q \) and \( I_Q^j \) respectively. At the end of this step, we have a list of pairs of visual synonym candidates \( T = \{ p_{r,t} \} \).

**Step 3: Visual synonym selection** In the third step, we acquire the final list of visual synonyms. We calculate the occurrence frequency \( f \) of all pairs of visual synonym candidates and we rank them accordingly. We then set a
3. EXPERIMENTAL SETUP

3.1 Implementation

Data set. We report our experiments on the Oxford5k data set, following the evaluation protocol suggested in [5].

Descriptors. We describe Hessian-Affine detected keypoints with SIFT. We use a 200K vocabulary, trained on the holiday data set [3].

Geometry estimation. We perform the geometric verification in the top $M = 30$ images. Very harsh RANSAC geometric constraints are imposed, requiring minimum $\gamma = 40$ inliers for accepting images $I_1, I_2$ as a positive match (threshold empirically found, data not shown). The maximum distance error for RANSAC is taken $\delta = 0.001$ and the approximation error $\epsilon = \delta/10$. In addition, we perform a spatial distribution inconsistency check [10].

3.2 Experiments

To assure that visual synonyms are not just the closest word pairs in descriptor space, we question in experiment 1:

- **Experiment 1: How close are visual synonyms in descriptor space?**

This experiment operates in the feature space, which in our case is the 128-D SIFT space. To answer this question, we calculate the distances between two synonym words $w^r$, $w^t$ and between the synonym words separately and the rest of the words of the vocabulary $w^j, j \neq r, t$. Since we have a vector space and we want to simply calculate vector distances, we use cosine similarity distance, that is $c(w^r, w^t) = \frac{\sum_i x^r_i x^t_i}{\|x^r\| \|x^t\|}$, where $x^r_i$ is the $i$-th coordinate of the feature vector of $w^r$.

In our second experiment we study the utility of visual synonyms for retrieval.

- **Experiment 2: Landmark image retrieval using visual synonyms**

We use visual synonyms in a visual augmentation framework, in order to enhance image representation. Our evaluation criterion for this retrieval experiment is the average precision score, which combines precision and recall into a single performance value.

4. RESULTS

4.1 Experiment 1: How close?

We show the results of experiment 1 in Figure 4. From the variety in words distance ranking, we conclude that visual synonyms are scattered throughout descriptor space. While some synonyms are relatively close neighbors indeed, the majority of synonym word pairs tend to be distant from each other. The results confirm that geometry links visual words that might indeed be far away in the descriptor space, no matter their common origins from the same physical elements in the 3D scenes.

4.2 Experiment 2: Landmark image retrieval

We show the results of experiment 2 in Figure 5. We consider as a baseline the standard bag-of-words model using a visual vocabulary of size 200K. Conform expectation, augmented retrieval using visual synonyms improves upon
the baseline. We obtain a MAP of 0.330 where the baseline achieves 0.305. When we compare augmented retrieval using visual synonyms with the state-of-the-art approach using complete vocabulary augmentation, we obtain a similar performance in terms of MAP (0.330 vs 0.325), yet we use on average 6 times less words. To be precise, we observe on average ∼4250 words in an image, all used during the full visual augmentation. On the contrary we observe on average only ∼700 of visual synonym words. The marginally better results, combined with the decreased number of words used, hint that meaningful words inside images were detected. An interesting case is the Pitt rivers scene, where complete visual augmentation performs best. As shown in Figure 6, for 3 out of 5 queries, amongst the retrieved results an image occurs with a signpost partly occluding the landmark. Occlusion affects more visual synonyms, since half of the potential synonyms hidden behind the signpost are missed. Nevertheless, we expect that visual synonyms can be further adapted and used to treat the occlusion element not as hindrance but as an appearance variation of the scene.

5. CONCLUSIONS

In this paper we have introduced the notion of visual synonyms, which are independent visual words that nonetheless have similar appearance. In order to detect them, we exploited the unchanged geometry of landmark images. We tested visual synonyms with respect to their closeness in the descriptor space. They have proven not to be simply the closest clusters, although they have similar appearance. Furthermore, visual synonyms have proven to achieve state-of-the-art performance in the bag-of-words retrieval pipeline, whilst using six times fewer visual words for augmenting query image histogram, as compared to complete visual augmentation.

6. REFERENCES