Feature Propagation on Image Webs for Enhanced Image Retrieval

Eric Brachmann
TU Dresden
01062 Dresden, Germany
eric.brachmann@tu-dresden.de

Marcel Spehr
TU Dresden
01062 Dresden, Germany
marcel.spehr@tu-dresden.de

Stefan Gumhold
TU Dresden
01062 Dresden, Germany
stefan.gumhold@tu-dresden.de

ABSTRACT

The bag-of-features model is often deployed in content-based image retrieval to measure image similarity. In cases where the visual appearance of semantically similar images differs largely, feature histograms mismatch and the model fails. We increase the robustness of feature histograms by automatically augmenting them with features of related images. We establish image relations by image web construction and adapt a label propagation scheme from the domain of semi-supervised learning for feature augmentation. While the benefit of feature augmentation has been shown before, our approach refrains from the use of semantic labels. Instead we show how to increase the performance of the bag-of-features model substantially on a completely unlabeled image corpus.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.4.6 [**Image Processing and Computer Vision**]: Segmentation

Keywords

content-based image retrieval; bag-of-features; image similarity; image webs; co-segmentation

1. INTRODUCTION

Measuring similarity between images is an essential part in many image processing applications. I.e. CBIR usually aims for retrieving images semantically similar to a specific query image. If no textual information is supplied one has to compute a visual similarity based on global or local image features.

This work focuses on the improvement of local image features. They are created by finding characteristic local patterns in images and describing them with (usually high-dimensional) feature vectors. The popular *bag-of-features* (BOF) [13] approach quantizes those vectors using a set of prototypical local patterns (also called *visual words*) and counting them in histograms. Two images are compared by measuring the intersection of their histograms. The more similar features they have in common, the more similar they are.

A common issue in feature based approaches for measuring image similarity are changes in acquisition conditions. The feature sets of images depicting identical objects vary under viewpoint changes until they stop to overlap. As the perspective shifts, features disappear and new features emerge. The robustness of common interest point descriptors can only partially compensate for this effect. The BOF similarity decreases until the model eventually fails even though the depicted object remains the same. Figure 1 illustrates the effect schematically.

We propose to overcome this issue by propagating features over a network of images. The method described in [4] can be used to construct image graphs of large, unstructured image collections. The construction itself is solely based on visual characteristics. No semantic knowledge is involved. Edges between images are established using *affine co-segmentation*. This web is used to propagate visual words among connected images using an adapted version of the method specified in [1].

Our results reveal that the transitive exchange of visual characteristics reduces the visual gap, caused by changing acquisition conditions. We benchmark our method on standard data sets, and show that through the additional robustness of BOF image signatures, retrieval performance in image search applications is substantially increased.

2. RELATED WORK

Sivic and Zisserman [13] introduced the bag-of-features image search in analogy to the bag-of-words search for text documents. Dictionaries of visual words are found by clustering feature descriptors of large, generic image collections. Thereupon, feature descriptors of arbitrary images can be assigned to the pre-calculated clusters which become visual words. Each image is represented by a histogram of visual words, a so called BOF. Feature locations and geometries are completely discarded. Entries of a visual word histogram $\mathbf{x}$ are weighted according to the *tf-idf* schema. Visual words that appear often in an image but are rare throughout the image collection receive a large weight. Similarity of two images is measured by the dot product $\mathbf{x_1} \cdot \mathbf{x_2}$ of their weighted visual word histograms. An inverted file structure facilitates fast image retrieval in very large image collections.

Since its initial proposal, many extensions of BOF search

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.

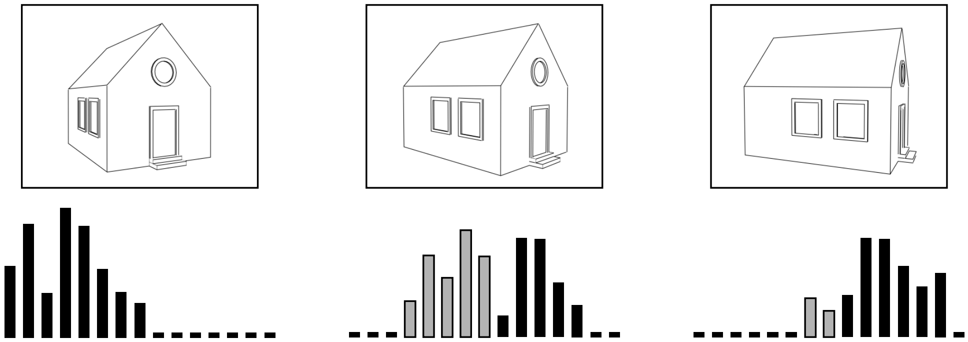

Figure 1: Illustration of BOF failure. Three images of the same object are depicted along with their visual word histograms. The viewpoint changes increasingly from left to right. The first and second histogram still overlap to large extents, marked in gray. However, the appearance of the object has changed too much in the third image. The overlap is minimal, and the similarity measure fails. Transitive feature exchange can prevent the failure.

systems have been published. With the resulting performance boost, BOF search systems still produce state of the art results on many image retrieval data sets[5]. Query expansion [3] takes the top retrieval results for an image, and re-runs the search by treating them as new queries. The approach is motivated by the observation that the top results are often relevant to the search query. The expanded query set is regarded as an enriched query representation. The retrieval results of all queries are combined, and ranked by similarity. By performing multiple searches for each query, query expansion multiplies retrieval times.

Spatial re-ranking[12] adds a verification step to BOF retrieval. It checks whether locations of matching features between the query and each top retrieval result are consistent by searching affine transformations between feature sets. Due to its computational cost, spatial re-ranking can only be applied to a small set of top retrieval results. Even so, it impairs the online response time of a retrieval system. We use the same approach of spatial verification to establish reliable image connections during image web construction. But in our case, it is done in a pre-processing step, that does not influence online query times.

Jegou et al.[5] present a complete state of the art BOF retrieval configuration. They augment BOF image signatures with binary strings that prevent wrong visual word matches even with coarse visual dictionaries. Instead of expensive spatial re-ranking after retrieval, they exploit weak geometric consistency (WGC). Simplified geometric information is embedded directly into the inverted file. It penalizes images during retrieval where matching features are inconsistent in terms of characteristic scale and dominant orientation compared to the query. A multiple assignment strategy prevents missing valid matches of similar features due to assignment to different visual words. The modifications of Jegou et al. require an adapted version of the inverted file structure with increased memory demand. Although cheaper than full spatial re-ranking, WGC constrains slow down retrieval. We adopt WGC constraints but instead of penalizing inconsistent images during retrieval, we use WGC checks during image web construction to further increase correctness probability when establishing image relations. The result of our approach is a set of enriched image signatures that may be used in any classical BOF retrieval system. The structure of the inverted file itself is not altered. In our experiments, we show that the size of the inverted file can be reduced by feature propagation. If desired, all above-mentioned modifications to BOF searches can be deployed along with our proposal.

Our approach is inspired by recent findings in [6]. The authors create visual and textual clusters of an image collection. The visual clustering is based on BOF similarity. The textual clustering is based on text labels. The text clusters thus represent (noisy) semantic information. The authors form extended visual clusters by combining the visual clusters with the textual clusters. Then, they distribute visual words in this extended visual cluster. This way, visual characteristics are exchanged between images that are semantically related. This benefits the retrieval performance.

We adopt the idea of [6], but go without any text labels or other semantic information. We deploy highly structured image relations in the form of image webs. In the following, we first describe our approach, state our experimental setup, and finally report the results we obtained after feature propagation.

3. IMAGE WEBS

Before we propagate features within an image collection, we discover its inherent image relations. A relation exists when two images display a common object. We largely follow the approach of Heath et al.[4] to construct highly interconnected image graphs that they coined image webs. Edge additions proceed in an order that leads to a fast rise in algebraic connectivity. This measure corresponds to the ability of a graph to distribute information. Hence, image webs are well suited for our purpose. Below, we summarize the image web construction briefly.

Heath et al. use affine co-segmentation to decide whether an edge exists between two images. Affine-invariant feature detectors locate salient image regions. Features are matched via their SIFT descriptors and reliable matches are selected using Lowe's ratio criterion[7]. A RANSAC-based, iterative process extracts subsets of matches that are related by an affine transformation between images. The union of feature subsets per image serves as a segmentation of the co-occurring image area. Heath et al. used these areas for visualization. We, however, are only interested in the fact whether at least one co-occurring area was found or not, and insert edges accordingly.

We extended the affine co-segmentation process in two aspects. Heath et al. used Harris affine[9], Hessian affine[9] and MSER[8] detectors to locate salient regions. We additionally included the ASIFT[14] detector for its robustness to viewpoint change. More affine co-segmentations succeed when it is included. ASIFTs large computation time and space requirements may pose a problem. We dealt with this issue by reducing image resolution for this detector.

Our second modification is an additional verification step for sets of co-segmented features. During our experiments, we observed a significant amount of wrong image associations, especially when the images showed repetitive patterns like nets, fences or texts. With such images, descriptor matches become arbitrary and chances are that some random subset adheres to an affine transformation. We accommodated for this by adapting the concept of weak geometric consistency (WGC)[5] constrains. For each matching feature pair found by affine co-segmentation we calculate the difference in characteristic scale and dominant orientation, respectively. Because these feature characteristics are computed in normalized local image frames the differences should be similar for corresponding sets of feature pairs, and diverse otherwise. WGC constraints were used by Jegou et al.[5] to re-rank retrieval results in BOF image searches. We deploy WGC to verify the validity of co-segmented feature pairs. We compute the variances of scale and orientation differences, $\sigma^2_{\Delta s}$ and $\sigma^2_{\Delta\alpha}$, of each co-segmented region. If the variances are larger than pre-determined thresholds, we deem the region inconsistent, and discard it.

This validation approach does not work with ASIFT features. Instead of affine normalization, it deploys affine simulations and calculates features in transformed image space. When sets of feature matches bridge different affine simulations, scale and orientation differences are inconsistent, even if the match is correct. Instead of excluding ASIFT from the WGC checks we calculate $\sigma^2_{\Delta s}$ and $\sigma^2_{\Delta\alpha}$ over the union of all detectors. Although ASIFT adds some distortion to these values, we are still able to define reliable thresholds to tell consistent and inconsistent feature match sets apart.

With these adaptions to the affine co-segmentation process we proceed with image web construction in two phases as Heath et al.[4] suggest: sparse web construction and densification. In the first phase, clusters of connected images are determined. A truncated BOF similarity ranking of image pairs of the corpus is formed. Affine co-segmentations are attempted in that order and edges are inserted where they succeed. No affine co-segmentation is performed for image pairs that already belong to the same connected component. This leads to a fast growth of sparsely connected image clusters. The first phase ends when the rate of successful co-segmentations drops below a threshold.

In the second phase, each cluster is augmented with edges that lead to a large increase in algebraic connectivity. Therefore, all remaining edges are ranked according to the absolute difference in the entries of the Fiedler vector associated with those images the edge would connect. The Fiedler vector is the Eigenvector corresponding to the second smallest Eigenvalue of a graphs Laplacian matrix. This Eigenvalue equals the algebraic connectivity. Affine co-segmentations proceed in the order of the new ranking until the algebraic connectivity of the current cluster converges. The image web construction is complete when all clusters were densified.

4. FEATURE PROPAGATION

We base our approach on *label propagation on similarity graphs* as presented in [1]. The authors discuss typical scenarios of semi-supervised learning where one has labeled and unlabeled data points. The goal is to spread known labels within a graph that covers the complete data set. This way, unlabeled samples receive existing labels from other samples.

The problem is broken down to one of class assignment. There are two classes: 1 and -1. The class affiliation is known for some samples, and unknown for others. In the latter case, the samples receive a class value of 0. This information is subsumed in a label vector $\hat{Y}$, that contains the initial class value for every node in the similarity graph. The similarity graph itself is represented by an affinity matrix W. The entries $W_{ij} \geq 0$ state whether the nodes i and j are related. The simplest variant is to set $W_{ij} = 1$ between connected nodes, and $W_{ij} = 0$ otherwise.

Once $\hat{Y}$ and W have been constructed, an iterative algorithm starts. In essence, the positive and negative class values of labeled samples influence the class values of unlabeled samples dependent on the local neighbourhood in the similarity graph. The procedure stops when the label vector $\hat{Y}$ converges. Finally, all samples whose entries in the label vector are negative receive class -1, and all samples whose entries are positive receive class 1. The authors of [1] describe two different algorithms: One where the classes of the labeled samples are fixed (they are reset after each iteration), and one where the classes of labeled samples may change during the procedure.

We regard every image cluster of an image web separately. In the following, we refer to these image clusters as image graphs. Every node in the graph is an image, the edges were established by affine co-segmentation. We construct the affinity matrix W dependent on a parameter k. The entry W_{ij} is 1 if the nodes i and j are connected by a path of length $k+1$ in the image graph, i.e. for $k = 0$, only direct neighbours are related in the affinity matrix W. I.e. k regulates the size of the local neighbourhood of an image. The diagonal entries $W_{ii} = 0$.

We regard each visual word as a separate label. The assignment of class values $\in \{-1, 0, 1\}$ is problematic. We only have the information whether a visual word appears in an image, or not. We cannot distinguish between labeled and unlabeled samples. If a visual word does not appear, we do not know whether it must not appear (class -1), or whether its appearance is unknown (class 0).

However, it is unreasonable to assign class values of 1 for visual word appearance, and class values of 0 for visual word absence. In that case, features would be distributed throughout the image graph until they reach all nodes. Negative class values are essential. Therefore, we set the class value to the appearance count $c \geq 1$ if a visual word appears in an image, and -1 otherwise. This way, we construct the label vector $\hat{Y}$ for all images of the image graph. We use the appearance count c instead of the class value 1 to increase the weight of visual words that appear multiple times in an image. Note that this variant does not involve unlabeled samples (class 0). Because of that, we cannot deploy the algorithm of [1] where the class assignments of labeled samples (classes -1 and 1) are fixed. Nothing would happen in our case. We use the variant, that allows initial classes to change.

Input for feature propagation are the affinity matrix W, the initial label vector $\hat{Y}^{(0)}$ depending on the current visual word, and a parameter $\alpha \in (0,1)$ that determines tendency of an image to keep its original signature. The smaller α the more difficult it is for the original signatures to change. Following [1], we construct a diagonal degree matrix D with

$$D_{ii} = \sum_j W_{ij}, \tag{1}$$

the sum of the rows of W. We also construct a diagonal matrix A,

$$A = \frac{\alpha}{1-\alpha}\ (D + \epsilon I) + I, \tag{2}$$

where I is the identity matrix and ϵ is a small term for numerical stability. Note that our equation for A is slightly simplified compared to the definition in [1] because all our samples are labeled. We proceed with the propagation as follows:

$$\hat{Y}^{(t+1)} = A^{-1}(\frac{\alpha}{1-\alpha}\ W\hat{Y}^{(t)} + \hat{Y}^{(0)}). \tag{3}$$

After the iteration converged, we assign the current visual word to all images i, where $\hat{Y}_i > 0$ with the appearance count $c = \lceil \hat{Y}_i \rceil$.

We repeat the whole process for every visual word in the dictionary. The set of new features for an image arises from all features with appearance counts $c > 0$ after propagation. Note, that originally existing features might disappear from an image if negative weights prevail in its local graph neighbourhood. As a result, it is possible that images end up without any features after propagation. We treat them like singular images outside the image web, and use their original signatures during retrieval.

We tested two different variants of incorporating the propagation results:

1. In the *default* variant, we substitute the original feature set of an image with the feature set after propagation.

2. In the *augmented* variant, we use the original feature set per image, and augment it with those features that were added during the propagation. I.e. the new signature is formed by the union of the original feature set with the feature set after propagation.

In an additional variant, we collected all visual words that disappeared from the image web during propagation. We speculated that these visual words, due to their irregular appearance, might be associated with clutter, and, therefore, harm retrieval. We erased them from the visual word dictionary and performed image search with the resulting filtered dictionary. The retrieval performance was clearly inferior to the baseline. Rigid exclusion of these visual words was harmful. We will not consider this approach any further.

5. EVALUATION

5.1 Datasets

We tested our approach on two data sets: INRIA holidays[5] and Oxford buildings[12].

Oxford buildings contains 5063 photos of several prominent buildings of Oxford along with some clutter images. Since certain objects are covered by many photos and some photos depict multiple objects, this data set is especially suited for image web construction. Because of its size, it represents a realistic application scenario. For each image, the authors provide pre-calculated features and pre-assigned visual words. Groundtruth consists of 55 queries with associated relevant images. The relevant images are divided into two groups, “good” and “ok”, depending on how much of the query object is visible. We do not differentiate between these two groups. An additional group “junk” consists of images that we ignore during evaluation as suggested in [12]. The data set refines queries with regions of interest which we do not use.

INRIA holidays contains 1491 personal holiday photos covering diverse natural scenes and man made environments. The structure of this data set differs considerably from Oxford buildings. It includes much more diverse, discontiguous scenes. Since only very few images belong together, it is much less suited for image web construction. Groundtruth is given in the form of 500 disjoint groups of related images. Each group contains only a small number of images, 3 on average. The first image of each group serves as query, and the remaining images are relevant retrieval results. Similar to the Oxford data set, the authors provide pre-calculated features for each image, but no pre-assigned visual words. Instead, they provide generic visual dictionaries ranging from 100 to $200k$ visual words. We assign visual words using FLANN[11] and the $200k$ dictionary. Furthermore, pre-calculated features of 1M random Flickr images are available on the INRIA holidays website. We use them to assemble distractor image signatures to test the robustness of our BOF implementation.

5.2 BOF Baseline

We implemented a basic BOF image search following the description of Sivic and Zisserman[13]. As has been suggested before[5], we deploy an adjusted *tf-idf* weighting. The original *term frequency* (*tf*) weight corresponds to a L_1 normalization of feature histograms. In our experiments, we achieved slightly better results with the L_2 norm. Similar to many retrieval scenarios we assess performance in terms of mean average precision (mAP). For computation of mAP, we adapted code published together with Oxford buildings. We use our implementation of the basic BOF image search to calculate baseline performance values on both data sets.

5.3 Web Construction

We used affine co-segmentation with the following parameters: We deploy the software of Mikolajczyk[10] to extract Hessian-Affine, Harris-Affine and MSER features. The ASIFT demo code[14] adds ASIFT features. In the case of ASIFT, we rescaled images by a factor of 0.4 to decrease the computational load. We perform feature matching with FLANN[11] and use Lowe’s ratio criterion[7] with $r = 0.7$ for match filtering. The RANSAC implementation of OpenCV[2] determines feature sets related by affine trans-

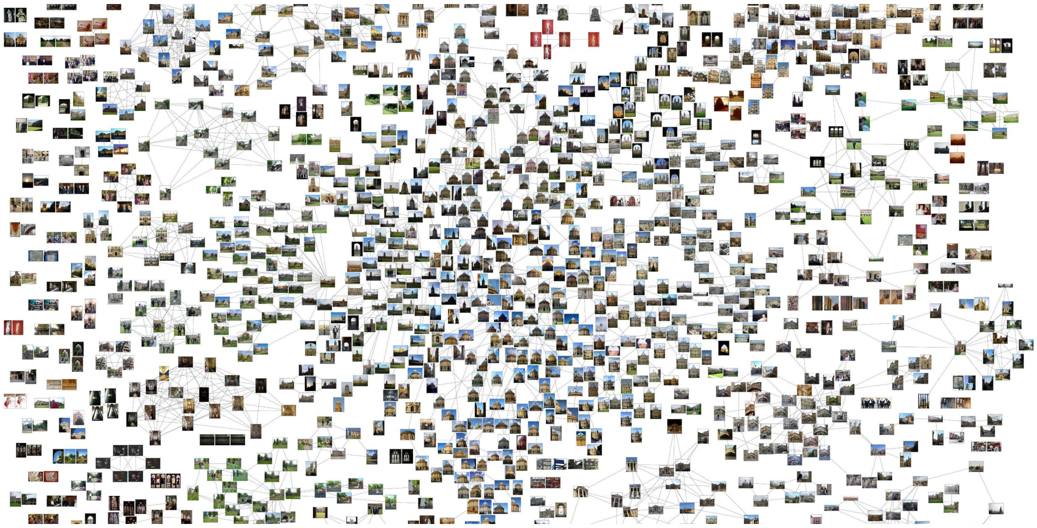

Figure 2: Part of the dense Oxford buildings image web. The largest cluster is clearly visible at the center. Smaller clusters are located towards the left and right margins.

formations with a reprojection error of 5 pixels. We accept feature sets if they consist of at least 20 features in both images.

For WGC checks after affine co-segmentation we allowed a maximal variance $\sigma^2_{\Delta\alpha}$ of 1.0 for orientation differences, and a maximal variance $\sigma^2_{\Delta s}$ of 0.1 for scale differences. We defined these thresholds after manually examining cases where affine co-segmentation yielded wrong results. We found that the variance of orientation differences is much more expressive than the variance of scale differences.

We tested our configuration of affine co-segmentation by manually validating its outcome on approximately 1300 image pairs of Oxford buildings. Only 13 of them were flawed.

Based on affine co-segmentation we constructed dense image webs of Oxford buildings and INRIA holidays. We stopped the initial sparse web construction when less than 20 co-segmentations were successful per 1000 image pairs processed. We stopped densification when the algebraic connectivity improved less than 5% of its initial rise. We found that a reasonable stopping criterion for densification is imperative. If all possible image connections are established, local image neighbourhoods become too big and generic for feature propagation. This results in decreased retrieval performance.

The Oxford buildings web consists of 363 distinct image clusters dominated by one large cluster with 547 images. The second largest cluster counts 100 images, and most of the cluster consist of 5 images or less. Altogether, ca 40% of the images appear in the image web. For all other images co-segmentation found no reliable partner. Reasons include the depiction of singular object, large changes in acquisition conditions, or image clutter. Figure 2 shows a part of the Oxford buildings web. The INRIA holidays web consists of 328 clusters with ca 50% of all images. All clusters are small with 2 to 10 images. Figure 3 shows one cluster of the INRIA holidays web in detail.

5.4 Propagation

Based on the image webs, we propagate features according to Section 4. Propagation depends on two parameters: α that determines the weight of the initial image signature. With a large α images tend to attract more features from their neighbourhood. The parameter k determines the size of the local image neighbourhood. During our preliminary experiments, we observed a value of $k = 1$ to be advantageous and fixed it for the experiments reported below. I.e., the local neighbourhood of an image consists of all images connected with a path of length 2. For α, we used values of 0.1, 0.5 and 0.9 to test strong, moderate and weak influence of the initial image signature.

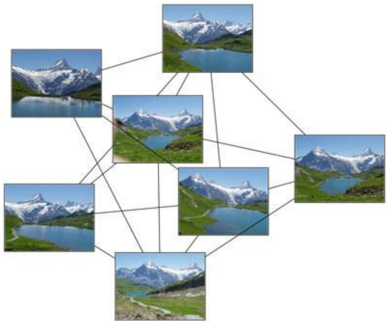

Figure 3: One image cluster of the INRIA holidays image web.

For images that do not appear in the image web, we keep the original signature. This also applies to images that end up without any features after *default* propagation. We compare mAP after propagation with the baseline values of the basic BOF search implementation. For the queries, we always use the original, unaltered image signatures when calculating the mAP.

We do not use distractor images during image web construction and feature propagation. We add distractor signatures afterwards to test the robustness of the feature propagation impact on retrieval performance.

5.5 Results

Table 1 subsumes our evaluation results on INRIA holidays used in conjunction with a dictionary of 200k visual words. With a basic BOF implementation we achieve a baseline mAP of 0.554 without distractor images. This is comparable to the baseline value reported in [5]. The results clearly show the benefit of feature propagation. *Default* feature propagation with $\alpha = 0.5$ results in a mAP of 0.594, i.e. an improvement of 7.1% over the baseline value. No propagation variant harms retrieval. We observe the benefit of a large α although there is no further improvement beyond $\alpha = 0.5$.

The impact of distractor images is straight forward. With more distractors added to the image collection, chances increase that they are confused with relevant images. MAPs are dropping for the baseline BOF search as well as for all propagation variants. However, the performance decrease is much smaller after feature propagation, see Figure 4. With 100,000 distractors the relative improvement over the basic BOF search rises to 30%. Image signatures clearly became more robust. Note that we used unaltered query signatures. Hence, mutual adaptions of queries and database images through feature propagation are ruled out.

For the most part, we can reproduce our observations for Oxford buildings. The baseline mAP is much smaller with 0.320. The data set contains more images of homogeneous objects, so there is more room for confusion. Furthermore, the homogeneous images exploit the expressiveness of the generic INRIA 200k dictionary only to some extent. Although performance is lower for Oxford buildings on absolute terms, the relative improvement through feature propagation is higher than for INRIA holidays. The best results are again achieved with *default* propagation and $\alpha = 0.5$. Without distactors, we boost the mAP to 0.409, an improvement of 27.7%. The improvment is stable in regard to distractors, see Figure 5. With 100,000 distractor images our best result is a mAP of 0.360 compared to the baseline of 0.223, a significant improvement of 60%.

We also performed feature propagation on the pre-assigned visual words of Oxford buildings. They are based on a much larger dictionary of 1M words that was furthermore learned on Oxford buildings itself. Naturally, it is much more expressive for this data set. We observe a high baseline mAP of 0.545. Here, we noticed dropping retrieval performance through feature propagation, see Table 3. With *default* propagation mAP drops by 9.3% for $\alpha = 0.5$, and by 3.3% for $\alpha = 0.9$. We attribute this to the sparseness of visual words with the 1M dictionary. Sparse visual words are more likely to vanish through *default* propagation. This can happen to an extent where the expressiveness of image signatures suffers. *Augmented* propagation prevents such effects. Indeed, with $\alpha = 0.5$ we achieve a mAP of 0.571, an improvement of 4.8%. We were not able to test the robustness with distractor signatures here, because the 1M word dictionary was not published.

5.6 Performance

The construction time of the web is dominated by the feature matching during affine co-segmentation. It took a few seconds each time on a single core (2.20 GHz). The sparse web construction has a complexity of $O(n)$ where n is the number of image pairs considered. We stopped after the rate of successful co-segmentations dropped below a threshold. Densification time depends on cluster size. In the worst case, all possible edges considered between images of an cluster are valid edges. Then, one potential edge is removed per iteration and all remaining potential edges have to be re-ranked. This results in a complexity of $O(p^2)$ for one cluster, where the number of potential edges $p = n \times k$ with n being the number of images in the cluster, and k being the number of most similar images considered to form image pairings. We used $k = 25$. For the larger Oxford buildings data set, image web construction took about a day on a single workstation.

The complexity of one propagation iteration is $O(n^2)$ in the number of images within the graph. The number of iterations depends on the convergence behavior of the label vectors, i.e. on the structure of the data set. The size of the dictionary determines the number of propagations that need to be performed. In our experiments, propagation times ranged from several minutes for INRIA holidays to several hours for Oxford buildings, again on a single workstation.

Our implementation does not exploit the various possibilities for parallelization at the later stages of feature propagation on image webs. Densification is independent for each cluster. Same goes for propagation for each visual word. Both processes can run in parallel, respectively. This does not apply for sparse web construction, and densification within one cluster because each co-segmentation does affect the list of subsequent co-segmentations.

Feature propagation has an effect on the amount of data a CBIR system has to manage. With *default* propagation, features from neighbouring images may be added to signatures, other features may vanish. In our experiments, the latter was the case far more often, and inverted files tended to become smaller. This is not possible in the *augmented* variant, where features can only be added to the initial image signatures. Thus, inverted files enlarge. In both variants, parameter α regulates the severity of the effect. A large α leads to a high decrease or increase of an inverted file, respectively. *Default propagation* decreases the inverted file size by 8% to 17% for INRIA holidays, and by 17% to 29% for Oxford buildings. *Augmentation* increases the inverted file size by up to 8% for INRIA holidays but only up to 3% for Oxford buildings.

Feature propagation is a one-time pre-processing step. It does not involve post-processing of retrieval results with additional operations, nor does it change the nature of image signatures. Thus, it does not influence query response times of CBIR systems compared to basic BOF retrieval. In theory, changing signature sizes could alter the accumulation time of similarity scores in the inverted file structure. However, we did not observe such an effect. Query times were stable.

6. CONCLUSION

Using affine co-segmentation, we constructed image webs of two data sets with several thousand images. We incorporated weak geometric consistency constraints in a novel way to estimate the reliability of a set of feature matches. This way, we detect and prevent wrong association between image pairs. As a result, we increase the correctness of the emerging image web considerably.

We used the image web to propagate visual words along image connections. To our knowledge, visual word propagation on completely unlabeled data is a new approach. We

Table 1: Evaluation of feature propagation on INRIA holidays in conjunction with a generic 200k word dictionary. The best performance per row is marked in bold face.

distractors	mAP baseline	mAP propagation *default*			mAP propagation *augmented*		
		$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
0	0.554	0.566	**0.594**	0.592	0.570	0.576	0.575
1,000	0.530	0.546	**0.576**	0.574	0.548	0.562	0.563
10,000	0.463	0.487	**0.533**	0.532	0.486	0.526	0.529
100,000	0.382	0.423	**0.498**	0.498	0.422	0.489	0.495

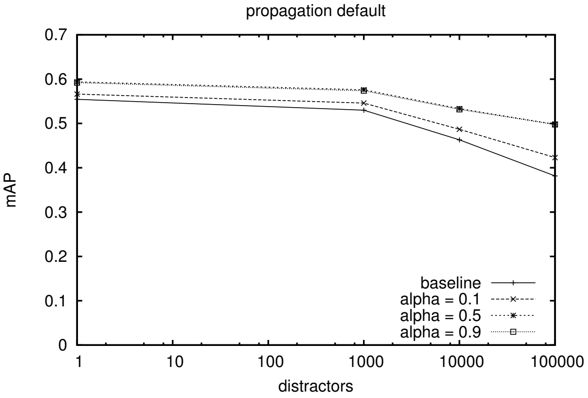

Figure 4: Impact of the number of distractor images on retrieval performance for INRIA holidays.

Table 2: Evaluation of feature propagation on Oxford buildings in conjunction with a generic 200k word dictionary. The best performance per row is marked in bold face.

distractors	mAP baseline	mAP propagation *default*			mAP propagation *augmented*		
		$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
0	0.320	0.370	**0.409**	0.338	0.334	0.365	0.368
1,000	0.315	0.366	**0.407**	0.338	0.327	0.359	0.363
10,000	0.294	0.345	**0.400**	0.333	0.306	0.340	0.344
100,000	0.223	0.284	**0.360**	0.309	0.235	0.271	0.284

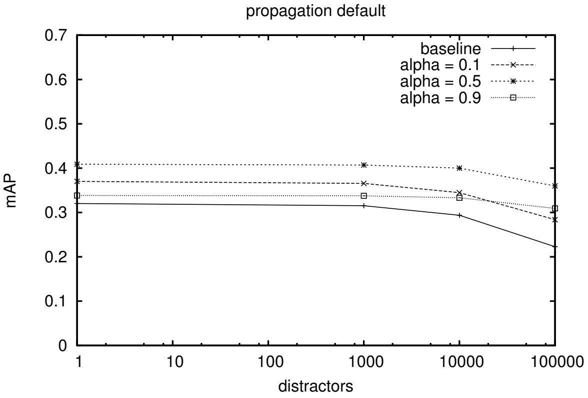

Figure 5: Impact of the number of distractor images on retrieval performance for Oxford buildings.

Table 3: Evaluation of feature propagation on Oxford buildings in conjunction with a data-set-specific 1M word dictionary. The best performance is marked in bold face.

mAP baseline	mAP propagation *default*			mAP propagation *augmented*		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.9$
0.545	0.553	0.494	0.527	0.553	**0.571**	0.564

showed, how techniques of semi-supervised learning can be adapted and deployed in this setup. We observed an increase in retrieval performance by up to 28%. This benefit is even more distinctive when distractor images are involved. Image signatures become considerably more robust. We also demonstrated that feature propagation is suited to reduce the amount of data necessary to describe an image collection. This is because sparsely distributed features are filtered out through propagation. This effect can cause problems for very large dictionaries that naturally entail sparsity. If too many features disappear retrieval performance suffers. This can be prevented when the original signatures stay fixed, and are only augmented through propagation.

The results of feature propagation can be easily incorporated into existing BOF search infrastructures. Feature propagation optimizes an inverted file without changing its composition. This separates it from approaches like Hamming embedding[5] or WGC constraints[5] that require a modified inverted file structure. Also, the propagation process runs in an offline stage, and does not impair the online query times of a system. The approach is complementary to other optimization strategies, like re-ranking based on spatial consistency[12], or query expansion[3]. It can thus boost the performance of state of the art BOF image searches even further.

Our feature propagation framework leaves many possibilities for variations. Image web connections can be weighted according to path length, image similarity or co-segmented area size. Feature exchange can be limited to features within co-segmented areas. This way, only sub-signatures of co-occurring objects would be enriched. Other methods than affine co-segmentation can be deployed, e.g. co-segmentation based on fitting homographies or fundamental matrices. Image relations could even be established based on full stereo reconstructions which yield further possibilities of verification. Fast web densification and feature propagation can be enforced by restricting the maximal cluster size during sparse web construction. A dissimilarity ranking could ensure sufficient heterogeneity for these small clusters. We leave these possibilities as future work.

7. ACKNOWLEDGMENTS

This work has partially been supported by the European Social Fund and the Federal State of Saxony within project VICCI (#100098171). This work was also supported by the DFG Priority Program 1335: Scalable Visual Analytics.

8. REFERENCES

[1] Y. Bengio, O. Delalleau, and N. Le Roux. Label Propagation and Quadratic Criterion. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 193–216. MIT Press, 2006.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

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

[4] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L. Guibas. Image webs: Computing and exploiting connectivity in image collections. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3432 –3439, june 2010.

[5] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[6] Y.-h. Kuo, H.-t. Lin, W.-h. Cheng, Y.-h. Yang, and W. H. Hsu. Unsupervised auxiliary visual words discovery for large-scale image object retrieval. *Discovery*, 1(c):1–8, 2011.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[8] J. Matas, O. Chum, U. Martin, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of British Machine Vision Conference*, volume 1, pages 384–393, London, 2002.

[9] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, Oct. 2004.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[11] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09)*, pages 331–340. INSTICC Press, 2009.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[13] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct. 2003.

[14] G. Yu and J.-M. Morel. ASIFT: An Algorithm for Fully Affine Invariant Comparison. *Image Processing On Line*, 2011, 2011.