Total Recall II: Query Expansion Revisited

Ondřej Chum Andrej Mikulík Michal Perdoch Jiří Matas

Centre for Machine Perception
Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague
[chum, mikulik, perdom1, matas]@cmp.felk.cvut.cz

Abstract

Most effective particular object and image retrieval approaches are based on the bag-of-words (BoW) model. All state-of-the-art retrieval results have been achieved by methods that include a query expansion that brings a significant boost in performance.

We introduce three extensions to automatic query expansion: (i) a method capable of preventing tf-idf failure caused by the presence of sets of correlated features (confusers), (ii) an improved spatial verification and re-ranking step that incrementally builds a statistical model of the query object and (iii) we learn relevant spatial context to boost retrieval performance.

The three improvements of query expansion were evaluated on standard Paris and Oxford datasets according to a standard protocol, and state-of-the-art results were achieved.

1. Introduction

Many successful particular object and image retrieval approaches are based on the bag-of-words (BoW) model introduced in [23]. In such retrieval methods, the query is represented by "a bag of" discretised descriptors of interest points called visual words. A shortlist of potentially relevant documents is efficiently retrieved by *tf-idf* scoring using inverted file. The candidate images in the shortlist are spatially verified. Finally, a new, "expanded" query, including features from verified shortlisted images, is issued.

Virtually all aspects of particular object BoW-type retrieval have been intensively studied: feature detectors and descriptors [14, 5, 25, 15, 16], vocabulary construction [23, 18, 21, 9, 17], spatial verification and re-ranking [21, 9], document metric learning [12, 10, 6] and dimensionality reduction [11, 20].

In this paper, we focus on the query expansion (QE) step. Automatic query expansion [7] has been shown to bring a significant boost in performance [7, 22, 10, 19], and all state-of-the-art retrieval results have been achieved by methods that include a QE step. Published QE methods focus on enriching the query model by adding spatially verified features. Retrieval with the "expanded" query follows. It has been observed that if the shortlist has enough true positives, the spatial verification re-ranking almost always correctly identifies relevant images, and, consequently, results for the expanded query are significantly better than the original single image query. Conversely, if BoW fails to the extent that there are no, or very few, correctly retrieved images in the shortlist, standard QE is of no help. One such situation, which arise in the presence of structures with multiple correlated features, have been referenced in the literature as

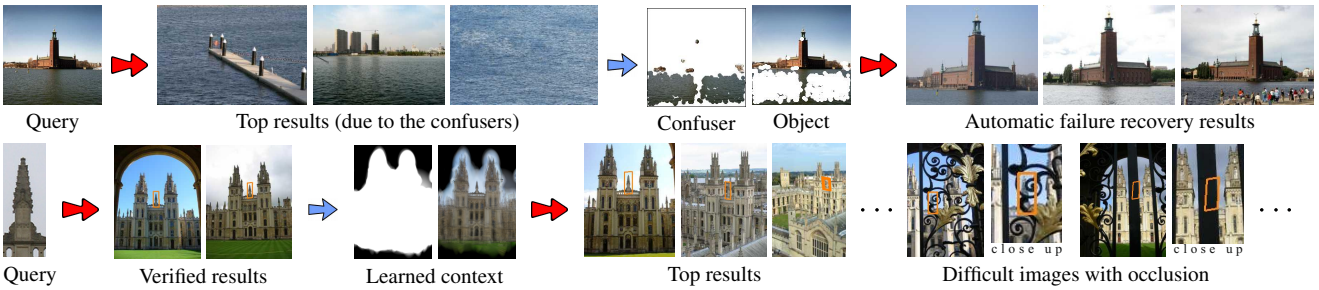


Figure 1. **Automatic Failure Recovery** (top): Initial retrieval results corrupted by confusing water features. The confuser model is learned dynamically. Successful subsequent query using the confuser model. **Context expansion** (bottom): Spatially consistent context is learned from retrieved images. This enables successful retrieval (before to first false positive image) and localization of heavily occluded objects.

cooc-sets [6] or confusers [13]. The *tf-idf* over-counting [6] caused by the correlated (second order statistics) structures cannot be alleviated by approaches analysing the first order (count) statistics of image features (*e.g*. [10]).

As the first contribution, we show how to detect and recover from the *tf-idf failure* situation. Unlike other approaches, the proposed method handles the presence of confusers in the query region on-the-fly, with no prior learning step required. The performance achieved is comparable to the state-of-the-art without the need for off-line and potentially time-consuming processing that is difficult to execute for a continuously updated database.

As the second contribution, we improve spatial verification and re-ranking by taking account of already evaluated results. The *incremental spatial re-ranking* (**iSP**) allows verification and subsequent use of images for query expansion that do not have a significant match against the original query, but do match a statistical model gradually built from the query and previously verified images.

As the third contribution, we propose a method that exploits spatial context by incorporating matching features outside the initial query boundary into the query expansion. Since the content outside the query region is not known at query time, the method requires efficient spatial verification of the retrieved images.

The rest of the paper is structured as follows: first, the components of BoW retrieval methods are reviewed in detail in Section 2, then the proposed automatic *tf-idf* failure recovery method is presented in Section 3. Finally, in Section 4, the novel incremental spatial verification and context growing are described and their performance is evaluated. Section 5 concludes the paper.

2. Baseline Query Expansion

Image representation. For each image in the dataset, affine invariant interest regions, called features, are detected. In this paper, a variant of multi-scale Hessian regions-of-interest [15] is used. For each feature, a 128-dimensional SIFT [14] descriptor is computed. Feature descriptors are vector quantized into visual words [23]. In this paper, the approach of approximate k-means with 1-million visual words was chosen from the many different vocabulary construction methods, such as [18, 21, 9].

Images are represented as collections of features, where each feature carries its visual appearance and spatial location. The visual appearance is captured as a visual word, while the spatial extent is encoded by an ellipse.

Bag of Words scoring. The search engines for particular object search have been inspired by widely used text search engines [2, 4]. The query and all the documents in the corpus are represented as a sparse vector of visual word occurrences. The search then proceeds by calculating the similarity between the query vector and each document vector. The standard tf-idf weighting scheme [3] is used, which deemphasizes the contribution to the relevance score from commonly occurring, less discriminative words.

For computational speed, the engine stores word occurrences in an index that maps individual words to the documents where the words are contained. For sparse queries, the use of an index, called inverted file, ensures that only documents that contain query words are examined, leading to a substantial speedup over the alternative of examining every document vector. The scores for each document are accumulated so that they are identical to the explicit computation of the similarity.

Spatial verification. As shown in [21, 19], the results can be significantly improved using the feature layout to verify the consistency of the retrieved images with the query region. The initial result list is re-ranked by estimating an affine transformation between the query image and result image. However, the spatial verification is significantly more time consuming than the BoW scoring, and is performed only on a shortlist of top scoring images. The shortlist is subsequently re-ranked based on the number of spatially verified inliers.

Query expansion is one of the standard methods for improving performance in text retrieval applications. A number of the highly ranked documents from the original query are re-issued as a new query. In this way, additional relevant terms can be added to the query.

In [7], the query expansion was introduced into the visual domain. A strong spatial constraint between the query image and each result enables an accurate verification, resulting in a suppression of false positives that typically ruin text-based query expansion. These verified images can be used to learn a probabilistic feature model to enable controlled construction of expanded queries.

In [7], a number of query expansion strategies were proposed. All of them follow a similar pattern: images in a shortlist are spatially verified against query features, images with sufficient numbers of matches (inliers) are back-projected by the estimated affine transformation into the query region, and, finally, a new query is issued. The differences in the proposed strategies are either in the number of repeated applications of the process, or in the method of feature selection.

The simplest well performing query expansion method is called average query expansion. A new query is constructed by averaging of document descriptors. This approach is the quickest from all the suggested strategies in [7], and also the most popular one [22, 19, 10]. We use the average query expansion as the baseline method.

3. Automatic tf-idf Failure Recovery in QE

The bag of words based scoring can be seen as a voting scheme. The necessary conditions for the voting scheme to work are that votes from features belonging to the query object are coincident in relevant images, and that background features randomly spread their votes among images in the database.

However, it has been observed in [8, 6, 13] that the assumption about noisy features does not always hold. Groups of correlated features typically occur on the water surface, on vegetation, images of text, faces, net-like structures, repetitive patterns, and statistical textures.

In the literature, the groups have been called cooc-sets [6], or confusers [13]. In the case where confusers appear in the query region, and are not related to the object of interest, the BoW retrieval often fails to select relevant images into the shortlist. This is a consequence of correlated voting for images that contain the same type of confusers, which suppresses the relative contribution of the specific object.

The proposed method is orthogonal to previously published approaches to QE, and can be used in conjunction with them. Previously, the methods of query expansion analysed results of the initial query to prepare a new query that is a *richer* representation of the object of interest. In contrast, the proposed approach tries to first obtain a *cleaner* model of the object by eliminating irrelevant confuser features.

We model the query (visual) words as a mixture of words generated by three processes (topics): the object words $\mathcal{O}$, the confuser words $\mathcal{C}$, and the random words $\mathcal{R}$. The three types of words, and their properties, are described in the following paragraphs.

We address the retrieval of *particular* objects, defined as a collection of features that preserves their appearance and spatial layout over a range of imagining conditions such as viewpoint change and scale change. The object words $w \in W_{\mathcal{O}}$ are likely to be observed in images containing the object of interest, *i.e.* $P(w|\mathcal{O})$ is high, $P(w|\mathcal{O}) \gg P(w)$. Moreover, the features associated with words, $w \in W_{\mathcal{O}}$, appear at fixed coordinates with respect to the canonical frame of the object, and thus allow for the geometric consistency check. The confuser words $w \in W_{\mathcal{C}}$ are defined as sets of correlated words, satisfying $P(w|\mathcal{C}) \gg P(w)$. However, confuser words are not significantly spatially consistent[1]. Randomly occurring words, $w \in W_{\mathcal{R}}$, generated from spurious features, and corrupted descriptors form the most frequently occurring class. As reported in [24], object features cover as few as 4% of the total features.

[1] We do not aim to solve a philosophical question regarding whether recurring objects, such as phone booths, are objects or confusers. According to our model, appearance and spatially consistent features form objects.

We propose a extension, called automatic *tf-idf* failure recovery, to the retrieval scheme. First, standard BoW retrieval with spatial verification is performed. The BoW scoring is used to produce a shortlist of documents. The images in the shortlist are checked for spatial consistency with the query features. The shortlist is significantly shorter than the size of the database[2]. If relevant images are included in the shortlist, they are identified by spatial verification and re-ranking. Once relevant documents are retrieved, automatic query expansion techniques are used to improve the object model $\mathcal{O}$. When a significant number of confuser words $\mathcal{C}$ is present in the query, the whole shortlist can be populated by images containing features generated from $\mathcal{C}$, and hence the spatial re-ranking cannot improve the search results. We call this situation a *tf-idf failure*. Even though the shortlist does not contain relevant images, it still conveys valuable information. A statistical model of the confusers $\mathcal{C}$ present in the query can be learned from the images in the shortlist, since a vast majority in the shortlist score higher than the relevant images. Once the confuser model $\mathcal{C}$ is known, its influence on the query is suppressed. There are three issues that need to be addressed: (i) efficiently estimate the confuser model $\mathcal{C}$, (ii) down-weight the effect of the confusers to the query result, and (iii) decide if the retrieval *tf-idf* failure has arisen.

Ad (i) The distribution $P(w|\mathcal{S})$ of visual words in the shortlist $\mathcal{S}$ is learned at virtually no cost during the tentative correspondence construction in the spatial re-ranking phase. Features whose visual words appear significantly more frequently than in the database are deemed to be part to the confuser model $\mathcal{C}$:

$$W_{\mathcal{C}} = \{w | P(w|\mathcal{S})/P(w) > r_0\}\,. \tag{1}$$

The likelihood ratio threshold was $r_0 = 10$ in our experiments.

Ad (ii) There are many options to reduce the influence of the estimated confusers $\mathcal{C}$. We choose to simply remove the confuser features from the query. This approach, while seeming naive, has been shown to be effective and efficient [6]. If a query expansion, average or any other type, is used after the failure recovery, features that back-project to regions occupied by the confusers are also removed. This prevents back-projected confuser features from entering the expanded query from the result images.

Ad (iii) To check whether a retrieval failure has arisen, we compare the estimated quality $\rho(Q)$ of results of two queries: the original query, Q_0, and the query after the recovery, Q_R. We estimate the quality of the results by the inlier ratios in the top matching results. First, the acceptable result images that each have an absolute and relative non-random number of inliers are selected. The score of the

[2] In our experiments, a shortlist of 1000 documents is used.

retrieval is then defined as the sum of inlier ratios over the acceptable results. Formally, let $\mathcal{S}_Q$ be a BoW shortlist of query Q, $T_Q(X)$ be a number of tentative correspondences between query Q and image X, and let $I_Q(X)$ be the number of geometrically consistent features between Q and X. The acceptable result of Q is a set of images

$$\mathcal{A}_Q = \left\{ X | X \in \mathcal{S}_Q \;\&\; I_Q(X) > I_0 \;\&\; \frac{I_Q(X)}{T_Q(X)} > \epsilon_0 \right\}.$$

The quality of the shortlist result of query Q is defined as

$$\rho(Q) = \sum_{X \in \mathcal{A}_Q} \frac{I_Q(X)}{T_Q(X)}. \quad (2)$$

To avoid wasted computation when improvement is unlikely, the estimated quality of the original query Q_0 is thresholded. If $\rho(Q_0) > \rho_0$, then the hypothesis of the *tf-idf* failure is directly rejected. In the experiments, the following parameters were used: minimal acceptable number of inliers $I_0 = 5$, minimal acceptable inlier ratio $\epsilon_0 = 0.2$, and the failure rejection threshold $\rho_0 = 5$. The algorithm is summarized in algorithm 1.

Algorithm 1 Automatic *tf -idf* failure recovery

Input: query features Q_0
Output: ⟨ query features, query results, feature mask ⟩

Execute query Q_0 including spatial verification
if $\rho(Q_0) > \rho_0$ **then** (see eqn. 2)
  return ⟨ Q_0, results (Q_0), empty ⟩
end if
Learn a set of confuser words $W_\mathcal{C}$ (eqn. 1)
$Q_R = Q_0 \setminus W_\mathcal{C}$
Execute query Q_R including spatial verification
if $\rho(Q_0) > \rho(Q_R)$ **then**
  return ⟨ return Q_0, results (Q_0), empty ⟩
else
  return ⟨ return Q_R, results (Q_R), mask($W_\mathcal{C}$) ⟩
end if

Efficiency. The proposed method introduces no extra cost for queries that return reasonable number of matching results (this is the case for almost all images in the standard Oxford and Paris datasets, where the query bounding box is tightly around the query building). For such queries the result is also unaffected because the original query is accepted. For other queries, one extra BoW scoring and spatial re-ranking step is executed. Since the new query is a subset of the original query, this additional step is faster than the original query.

3.1. Experimental Results

In this section, we compare the results of the confuser model learned in the proposed automatic *tf-idf* failure recovery step with results obtained by cooc-sets [6]. The quantitative results on the Q8 dataset [6] embedded in a database of over 5 million images are shown in Tab. 1. It is not feasible to obtain all true positives, so the average precision $\widehat{\text{AP}}$ is only an upper bound estimate. New positive results have been discovered by the proposed method, and the $\widehat{\text{AP}}$ values are not directly comparable to values in [6].

	AFR		Cooc [6]		Baseline	
	$\widehat{\text{AP}}$	fFP	$\widehat{\text{AP}}$	fFP	$\widehat{\text{AP}}$	fFP
Stockholm	**0.659**	**16**	0.569	15	0.032	1
Dragon Wall	**0.797**	**56**	0.726	52	0.065	5
St Ignazio	**0.945**	**17**	0.737	14	0.105	2
Colloseum	**0.762**	**514**	0.136	85	0.018	13
Barcelona	**0.895**	**17**	0.789	15	0.053	1
St Mary	**0.943**	**57**	0.895	51	0.020	1
Vaticane	**0.957**	**22**	0.870	20	0.130	3
Bridge	0.583	4	**0.716**	**5**	0.143	1

Table 1. Quantitative comparison of the proposed method with [6] and the baseline method on the Q8 dataset: Estimated average precision $\widehat{\text{AP}}$ and the rank of the first false positive fFP.

Table 1 shows, that for most cases, the two methods give comparable results. For the 'Colloseum' query, the proposed approach gives significantly better results than results obtained by the cooc-sets approach. This is because the cooc-sets approach excludes object-relevant cooc-set features as well as the confuser features, as opposed to the proposed approach which correctly learns the confuser model.

Pros and Cons of the proposed method. **Pros:** No prelearning step is required, so the method is applicable to any dataset and any vocabulary, and additionally, it does not require good training sets that generalize well, or retraining for different vocabularies. The method is specific to the current database and to the current query, so features for some queries that are confusers can be useful for other queries. **Cons:** The proposed method requires the execution of the original query, while the cooc-set approach can filter confuser features beforehand. In some queries, the confuser features may represent significant proportion of the features and thus the full query takes longer to execute.

4. Improving Blind Relevance Feedback in QE

In QE, spatial verification and re-ranking plays the role of blind relevance feedback. Spatially consistent images retrieved with the original query are deemed "relevant", similarly to the images chosen by the user in manual relevance feedback. The selected parts of "relevant" images then contribute to the new, expanded query. The quality of the decision on relevance significantly influences the success of query expansion.

In this section, two improvements of spatial re-ranking are presented. First, we introduce *incremental spatial re-*

ranking (**iSP**), where the verification accounts for not just spatial agreement with the initial query, but also agreement with all previously verified images. Second, we show that it is beneficial to "grow" the model of the object beyond the boundaries of the initial query, and to examine the spatially consistent neighbourhood of the query.

4.1. Incremental Spatial Re-ranking

In this section, an improvement of the spatial re-ranking (**SP**) phase of the baseline method (see Section 2) is proposed. As in the baseline method, the novel incremental spatial re-ranking (**iSP**) starts with the shortlist $\mathcal{S}$ of images ordered by the BoW score. The objective of **iSP** is to form a statistical model of the query object.

Initially, the statistical model M^0 includes only features from the query. Next, images in the shortlist are considered in the order given by BoW scoring. Each image $X \in \mathcal{S}$ is geometrically matched against the current model M^i. If the image matching quality $I_{M^i}(X)$ is greater than θ, the query object model is updated, and M^{i+1} is formed.

The quality function $I_{M^i}(.)$ is defined as the number of geometrically consistent features with the same visual word in image X and model M^i. The threshold θ was set to 15 after extensive preliminary experiments. The updated model M^{i+1} is the union of features in model M^i and features in image X, back-projected using function $f(.)$ onto the query image, clipped by the query bounding box. The final ranking of a shortlisted image is defined by the quality function. The method is described in Algorithm 2.

Since the simplest quality measure described above performed well, no alternatives, *e.g.* accounting for inlier ratio, geometric overlap, or weights of matching features, were evaluated.

Algorithm 2 Incremental spatial re-ranking

Input: query image X_q, shortlist $\mathcal{S}$ of images
Output: ranking $R : \mathcal{S} \leftrightarrow \{1..|\mathcal{S}|\}$, expanded model M^n of the object

$M^0 := X_q$
$Q := [], i := 0$ \\ $Q[k]$ records the number of inliers
for $k := 1$ to $|\mathcal{S}|$ **do**
  $X := \mathcal{S}[k]$
  $Q[k] := I_{M^i}(X)$
  if $Q[k] > \theta$ **then**
    $M^{i+1} := M^i \cup f(X)$
    $i := i + 1$
  end if
end for
$R :=$ ranking of the images according to $Q[k]$.

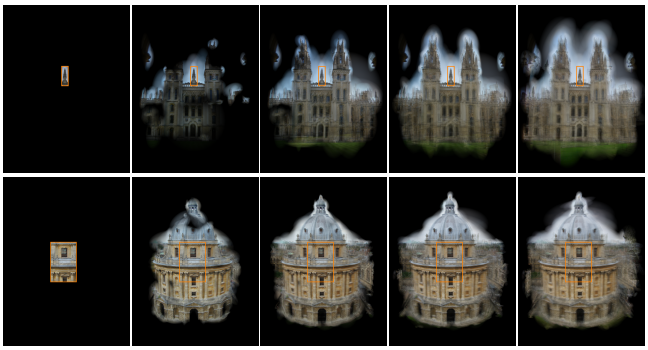


Figure 2. The process of context learning. Left column: the original query. Other columns: feature patches back-projected into the spatial context from 2, 5, 10 and 20 spatially verified images.

4.2. Outside the Query Boundaries: Incorporating Spatial Context

The content outside the query region is not known at the query time. It is clear that learning the query context must be done by the "matching results to results" approach. The process of the *spatial context learning* takes place either after spatial re-ranking, or, in the case of **iSP**, after each update of the query object model. The latter has the advantage that an image may be verified with the help of the context. In this case, implementation of context growing is trivial. As in **iSP**, features are back-projected to query image and are added to the model regardless of whether they are inside or outside the query bounding-box. The extension of the object model beyond the boundary of the original query only requires relaxing this constraint.

At the beginning of the learning phase, the spatial context is identified with the area inside the query boundary. A feature added into the model that is not inside the context is inactive until confirmed by feature(s) from another image with the same visual word and similar geometry. Once a feature is confirmed, it adds the neighbourhood around its center to the context. All confirmed features in the context are treated as active. The active features are considered the same as those inside the bounding box, and are used in spatial verifications, and, finally, in the query expansion. This is efficiently implemented by spatial binning. The process is summarized in Fig. 1 (bottom).

The progress of the spatial context growth for two queries is visualized in Fig. 2. The learned model of the query is shown as the mean of elliptic patches associated with its features back-projected to the query. The query bounding box is drawn as an orange rectangle. To save space, the area not covered by the model, or equivalently, the area not covered by a single feature, is cropped. Experiment 2, summarized in Tab. 4, shows that including the spatial context improves performance.

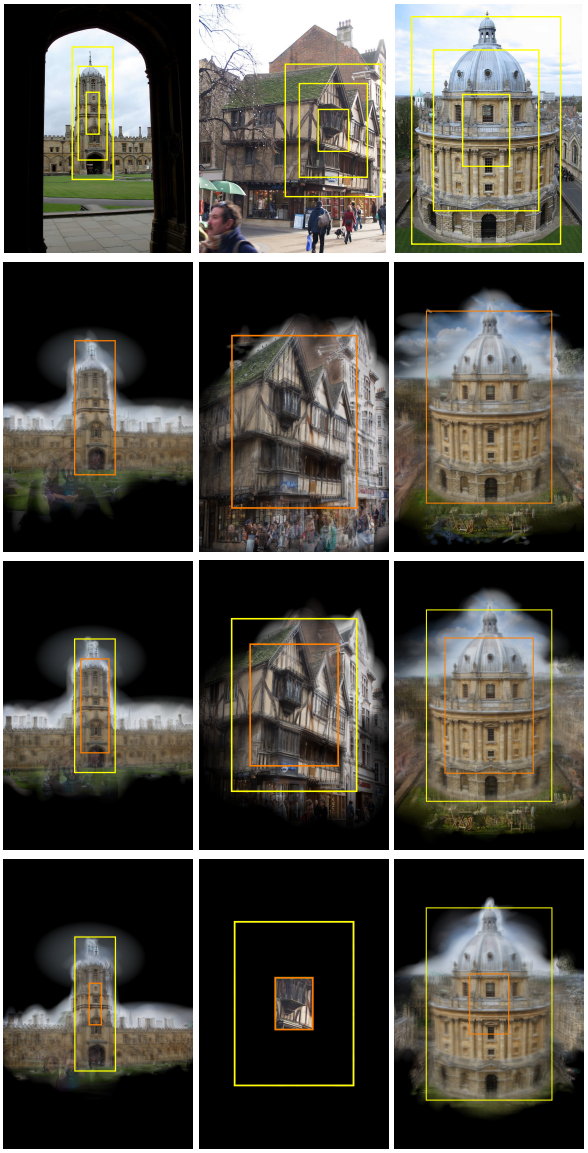


Figure 3. Top row: Examples of the full (100%) bounding box of some Oxford protocol queries (outer rectangle) and the query bounding boxes reduced to 50% and 10%. Rows 2,3 and 4 depict the context learned from the full, 50% and 10% bounding boxes respectively (the orange rectangles). The yellow rectangle shows the original bounding box. Note the ability of the **iSP + ctx QE** to learn the context even from the smallest query. The method failed on the CORNMARKET 10% (bottom row, middle) due to the insufficient number of spatially verified images.

4.3. Experiments

Datasets and Evaluated Methods

The image retrieval methods proposed in Sections 4.1 and 4.2 were evaluated according to the standard protocol [21] on the Oxford Buildings [21] and Paris datasets [22]. Additionally, ≈100k confuser images of the 75 most popular Flickr tags are provided with the Oxford Buildings dataset [21]. The datasets used in experiments are presented in Tab. 2. For each of the Oxford and Paris datasets, the evaluation protocol defines 55 queries, five for each landmark, with precise ground truth. The performance of all retrieval experiments is measured using the mean average precision (mAP), *i.e.* the area under the precision-recall curve for the further details see [21].

Oxford 5k	5062 images of 11 Oxford landmarks
Oxford 105k	Oxford 5k + 100k Flickr distractors
Paris 6k	6414 images of 11 Paris landmarks
Paris 106k	Paris 6k + 100k Flickr distractors

Table 2. The datasets used in experiments.

An extended protocol over the Oxford and Paris datasets is introduced. New, smaller query bounding boxes covering 10% to 90% of the original bounding box were introduced to evaluate the image retrieval approaches in more challenging conditions. The cropped bounding boxes and the ground truth are available through [1].

The proposed incremental spatial re-ranking and spatial context growing methods are compared with the state-of-the-art image retrieval approaches, see the list in Tab. 3. All methods use, in experiments on the Oxford dataset, a 1M visual word vocabulary trained on the Paris dataset and vice versa.

1	**SP**	BoW scoring, spatial re-ranking, no query expansion, see Section 2
2	**iSP**	BoW scoring, incremental spatial re-ranking, no query expansion
3	**SP + avg QE**	BoW scoring, spatial re-ranking, average query expansion, see Section 2
4	**iSP + avg QE**	BoW scoring, incremental spatial re-ranking, average query expansion
5	**SP + ctx QE**	BoW scoring, spatial re-ranking, context query expansion, see Section 4.2
6	**iSP + ctx QE**	BoW scoring, spatial re-ranking, incremental spatial re-ranking with context and context query expansion.

Table 3. Description of the state-of-the-art (rows 1 and 3) and the proposed methods (rows 2,4,5 and 6).

Experiment 1. Evaluation of Incremental Spatial Re-ranking

The experiment compares all image retrieval methods listed in Tab. 3 on the Oxford and Paris datasets. We observe that **iSP** outperforms **SP** in all cases; compare the left and right columns of sections I, II and III of Tab. 4. The **iSP** improves performance by approximately one half of the query

expansion effect; compare columns I right, and II left. Since only the shortlist is accessed, the performance improvement is obtained at a negligible cost compared to issuing a second query. This encourages the use of **iSP** instead of the standard **SP** re-ranking. Additionally, the benefits of **iSP** and query expansion are additive; compare columns I right and II right. Finally, adding spatial context has negligible effect on the Oxford dataset and improves performance on the Paris dataset. This is due to the fact that on the Oxford protocol, queries include entire objects, and there is little gained by growing the spatial context.

	I. w/o QE		**II. avg QE**		**III. ctx QE**	
	SP	**iSP**	**SP**	**iSP**	**SP**	**iSP**
Oxford 5k	0.616	0.741	0.785	0.825	0.781	0.827
Oxford 105k	0.553	0.649	0.725	0.761	0.731	0.767
Paris 6k	0.617	0.679	0.720	0.772	0.753	0.805
Paris 106k	0.508	0.556	0.627	0.687	0.653	0.710

Table 4. Comparison of image retrieval methods with standard (**SP**) and incremental spatial re-ranking (**iSP**).

Experiment 2. Evaluation of Spatial Context Expansion

Next we study the influence of incorporating the spatial context of the query *i.e*. extending the model of the query outside its bounding box. The behaviour is demonstrated on the same datasets by using a novel protocol.

As shown in experiment 1, the effect of context learning is not significant in the case of the Oxford dataset. To model a situation where only a detailed or partial view of the object is available, the following protocol was devised: The query bounding boxes were symmetrically reduced to 10% of their area in nine steps, see Fig. 3. The maximum spatial extent of the context was limited to an area 25$\times$ larger than the reduced query bounding box.

The results (see Fig. 4) show that the performance of the retrieval method using both spatial context and incremental spatial re-ranking (**iSP + ctx QE**) drops below the state-of-the-art (black dashed line in Fig. 4) method only after reducing the bounding box area to 40%, (Fig. 4b,d), or even to 20% (Fig. 4a,c) of the full query bounding box. One of the reasons for the drop in performance is that to keep the number of features in the model, and thus the speed of spatial re-ranking reasonable, we limit the number of images added to the model to ten, which is insufficient to reconstruct the model to the quality of the original query. Also, the results of initial queries on the standard datasets already contain many of true positives, and even the standard query expansion manages to retain a sufficient model of the object.

Some examples of spatial contexts learned for some of the Oxford protocol queries are shown in Fig. 3.

5. Conclusions

We have proposed three extensions of the query expansion step in the BoW-based particular object and image retrieval. First, a method capable of preventing *tf-idf failure* caused by the presence of confusers was introduced. Second, the spatial verification and re-ranking step was improved by incrementally building a statistical model of the query object. Finally, we show that relevant spatial context improves retrieval performance.

The proposed improvements of query expansion were evaluated on established Paris and Oxford datasets according to a standard protocol and state-of-the-art results were achieved.

Finally, a new and more challenging protocol over standard datasets was introduced [1].

Acknowledgements. The research was supported by Czech Government research program MSM6840770038, CTU SGS11/125/OHK3/2T/13, EC project FP7-ICT-247022 MASH, and a Microsoft scholarship.

References

[1] http://cmp.felk.cvut.cz/data/tr2/, www.

[2] Y. Aasheim, M. Lidal, and K. Risvik. Multi-tier architecture for web search engines. *Web Congress, 2003. Proceedings. First Latin American*, 2003.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, ISBN: 020139829, 1999.

[4] L. Barroso, J. Dean, and U. Holzle. Web search for a planet: The google cluster architecture. *Micro, IEEE*, 23, 2003.

[5] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, 2006.

[6] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *Proc. CVPR*, 2010.

[7] 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 *Proc. ICCV*, 2007.

[8] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

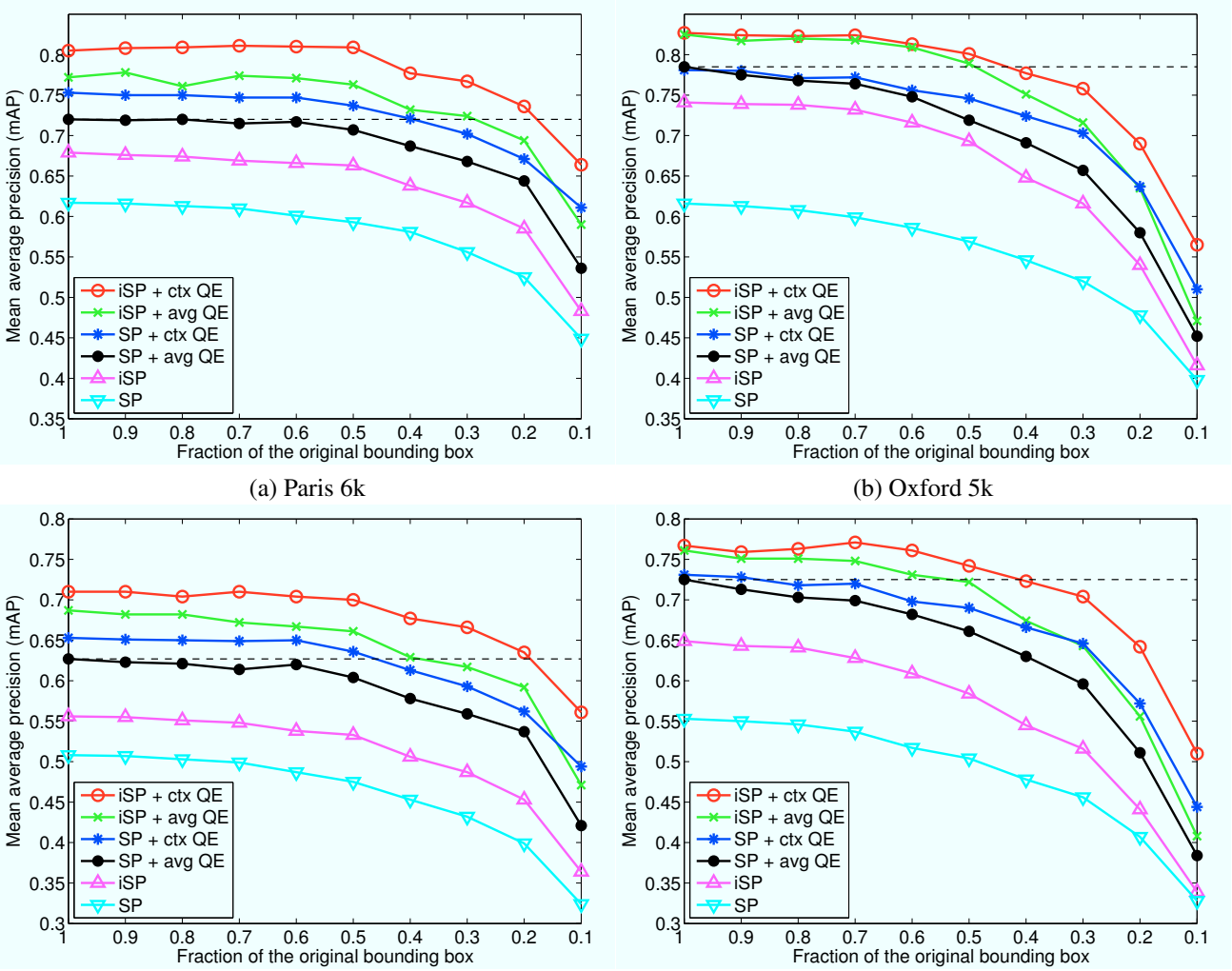
[9] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008.

[10] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *Proc. CVPR*, 2009.

[11] H. Jégou, M. Douze, C. Schmid, and P. Prez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010.

[12] H. Jégou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Proc. CVPR*, 2007.

[13] J. Knopp, J. Sivic, and T. Pajdla. Avoding confusing features in place recognition. In *Proc. ECCV*, 2010.



(a) Paris 6k

(b) Oxford 5k

(c) Paris 106k

(d) Oxford 105k

Figure 4. The influence of decreasing the query bounding box size on image retrieval methods. The black dashed line is the performance of the state-of-the-art [7] method with the original bounding box. The performance of the proposed **iSP + ctx QE** is superior to the state-of-the-art method, if the query covers more than 20% of the bounding box on the Paris datasets, and more than 40% of the bounding box on the Oxford datasets. The compared methods are listed in Tab. 3.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[15] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 1(60):63–86, 2004.

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

[17] A. Mikulik, M. Perdoch, O. Chum, and J. Matas. Learning a fine vocabulary. In *Proc. ECCV*, pages 1–14. Springer, 2010.

[18] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006.

[19] M. Perdoch, O. Chum, and J. Matas. Efficient representation of local geometry for large scale object retrieval. In *CVPR*, 2009.

[20] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proc. CVPR*, 2010.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[22] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. CVPR*, 2008.

[23] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003.

[24] P. Turcot and D. G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *ICCV Workshop LAVD*, 2009.

[25] S. Winder, G. Hua, and M. Brown. Picking the best daisy. In *Proc. CVPR*, 2009.