Multiple queries for large scale specific object retrieval

Relja Arandjelović
relja@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Department of Engineering Science
University of Oxford
Parks Road
Oxford, OX1 3PJ, UK

The aim of large scale specific-object image retrieval systems is to instantaneously find images that contain the query object in the image database. Current systems, for example Google Goggles, concentrate on querying using a single view of an object, e.g. a photo a user takes with his mobile phone, in order to answer the question "what is this?". Here we consider the somewhat converse problem of finding *all* images of an object given that the user knows what he is looking for; so the input modality is text, not an image. This problem is useful in a number of settings, e.g. media production teams are interested in searching internal databases for images or video footage to accompany news reports and newspaper articles.

Given a textual query (e.g. "Fontana di Trevi"), our approach is to fetch images of the queried object using textual Google image search. These images are used to visually query the database to discover images containing the object of interest. We compare a number of methods for combining the multiple query images, including discriminative learning. We show that issuing multiple queries significantly improves recall and enables the system to find quite challenging occurrences of the queried object. Fig. 1 shows an example of this process, which proceeds in a matter of seconds from typing the query to receiving the ranked results.

Using multiple queries overcomes a number of the shortcomings of existing large scale specific object retrieval methods. It is important to first consider why images containing the target object are missed using a single query. Addressing this problem has been one of the main research themes in specific object retrieval research with developments in feature encoding to alleviate vector quantization (VQ) losses [4, 5, 6], and in augmentation of the bag of visual word (BoW) representation to alleviate detector and descriptor drop out (as well as, again, VQ losses) [1, 2, 3].

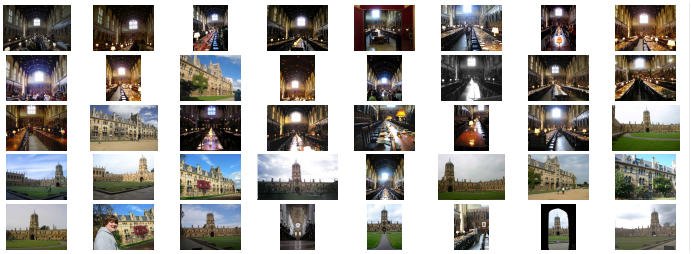
The limitation of current augmentation approaches, which are based on query expansion (QE) within the data set, is that they rely on the query to yield a sufficient number of high precision results in the first place. In more detail, in QE an initial query is issued, using only the query image, and confident matches, obtained by spatial verification, are used to re-query. There are three problems with this approach: (i) It is impossible to gain from QE if the initial query fails. (ii) If the dataset does not contain many images of the queried object QE cannot boost performance. (iii) It is not possible to obtain images from different views of the object as these are never retrieved using the initial query, for example querying using an image of a building façade will never yield results of its interior.

Table 1 shows the retrieval performance on the Oxford 105k dataset. It can be seen that all the multiple query methods are superior to the "single query" baseline, improving the performance by 29% and 52% for the Oxford queries and Google queries (with spatial reranking), respectively. It is clear that using multiple queries is indeed very beneficial as the best performance using Oxford queries (0.937) is better than the best reported result using a single query (0.891 achieved by [1]); it is even better than the state-of-the-art on a much easier Oxford 5k dataset ([1]: 0.929). All the multiple query methods also beat the "best single query" method which uses ground truth to determine which one of the images from the query set is best to be used to issue a single-query.

	Google queries		Oxford queries	
	W/o SR	With SR	W/o SR	With SR
Single query	0.464	0.575	0.622	0.725
Best single query ("cheating")	0.720	0.792	0.791	0.864
Joint-Avg	0.834	0.873	0.886	0.933
Joint-SVM	0.839	0.875	0.886	0.926
MQ-Max	0.746	0.850	0.826	0.929
MQ-Avg	0.834	0.868	0.888	0.937
MQ-ESVM	N/A	0.846	N/A	0.922

Table 1: **Retrieval performance (mAP) of the proposed methods on the Oxford 105k dataset.** SR stands for spatial reranking. The "Oxford queries" (OQ) and "Google queries" (GQ) columns indicate the source of query images, the former being the 5 predefined query images and the latter being top 8 Google images which contain the queried object. All proposed methods significantly outperform the "single query" baseline, as well as the artificially boosted "best single query" baseline.

(a) Top 8 Google Image results for the textual query "Christ Church, Oxford"



(b) Top 40 retrieved results from the Oxford 5k dataset by searching with the Google images

Figure 1: **Multiple query retrieval.** Images downloaded from Google using the "Christ Church, Oxford" textual query (a) are used to retrieve images of Christ Church college in the Oxford Buildings dataset (b). All the top 40 results of (b) do show various images of Christ Church (the dining hall, tourist entrance, cathedral and Tom tower). This illustrates the benefit of issuing multiple queries in order to retrieve all images of the queried object.

Figure 2: **Retrieved images from the TrecVid 2011 dataset.** The textual queries used to download images from Google and use them to retrieve images from the TrecVid 2011 KIS dataset are: (a) Presidential seal, (b) EA sports logo, (c) Leonardo da Vinci, (d) Fontana di Trevi, (e) I want you, (f) Comedy central logo.

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012.

[2] 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.

[3] O. Chum, A. Mikulik, M. Perd'och, and J. Matas. Total recall II: Query expansion revisited. In *Proc. CVPR*, 2011.

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

[5] A. Mikulik, M. Perd'och, O. Chum, and J. Matas. Learning a fine vocabulary. In *ECCV*, 2010.

[6] 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.