Abstract

In this work, we address the issue of geometric verification, with a focus on modeling large-scale landmark image collections gathered from the internet. In particular, we show that we can compute and learn descriptive statistics pertaining to the image collection by leveraging information that arises as a by-product of the matching and verification stages. Our approach is based on the intuition that validating numerous image pairs of the same geometric scene structures quickly reveals useful information about two aspects of the image collection: (a) the reliability of individual visual words and (b) the appearance of landmarks in the image collection. Both of these sources of information can then be used to drive any subsequent processing, thus allowing the system to bootstrap itself. The main result of this work is that this unsupervised “learning-as-you-go” approach significantly improves performance; our experiments demonstrate significant improvements in efficiency and completeness over standard techniques.

1. Introduction

In this work, we address the issue of geometric verification, with a focus on modeling large-scale landmark image collections. In particular, we show that we can compute and learn descriptive statistics pertaining to the image collection by leveraging information that arises as a by-product of the matching and verification stages. In designing a 3D reconstruction system for internet photo collections, one of the key considerations is robustness to “clutter” – when operating on datasets downloaded using keyword searches on community photo sharing websites (such as Flickr), it has been observed that invariably, a large fraction of images in the collection are unsuitable for the purposes of 3D reconstruction [4, 6]. Thus, one of the fundamental steps in a 3D reconstruction system is geometric verification: the process of determining which images in an internet photo collection are geometrically related to each other. This is a computationally expensive process, and much work in recent years has focused on developing efficient ways to perform this step. For example, Agarwal et al. [1] use image retrieval techniques to determine, for every image in the dataset, a small set of candidate images to match against. An alternate approach, adopted by Frahm et al. [3], is to first cluster the images based on global image descriptors and to then perform the verification within each cluster. While these approaches are extremely promising, there are still some limitations. For instance, even the carefully optimized approach described in [3] suffers from “incompleteness”; due to the coarse clustering, a large fraction of images are discarded following the clustering and verification steps. In this work, we aim to overcome these limitations.

2. Approach

Thus far, the typical way to perform geometric verification has been to estimate the geometric relationship between pairs independently. In other words, given a pair of images, features are matched between the images to obtain a set of putative correspondences, and then a robust estimation algorithm is used to identify a set of inliers. This process then repeats for the next pair of images, typically ignoring the results produced by any previous rounds of verification. Our main idea in this work is simple: as the geometric verification progresses, we learn information about the image collection, and subsequently use this learned information to improve efficiency and completeness. More specifically, since images of the same geometric structures are being repeatedly verified against each other, this process of repeated matching reveals useful information about two things:

(a) the stability and validity of low-level image features

(b) the global appearance of the landmarks in the dataset

While current techniques either ignore this information, or leverage it for other tasks via an offline processing stage, we feed this information directly back into the verification
the absence of any prior information, we assign that is proportional to the validity of the visual word. In addition, consider a set of visual word weighting tends to emphasize visual words that are stable geometric check in other image pairs (see Figure 1(b)). This image by the number of times it has previously passed the same scene, we can weight each visual word in the current pipeline. This approach, while simple, is also very effective; our results demonstrate significant improvements in efficiency compared to current techniques.

2.1. Identifying useful visual words

As a motivating example, consider Figure 1(a), which shows all detected SIFT features for a single image. Note that a large number of features lie in areas of the image that are very unlikely to pass any geometric consistency check (for e.g., features on vegetation, people, and in the sky). Now, if we have previously verified other images of the same scene, we can weight each visual word in the current image by the number of times it has previously passed the geometric check in other image pairs (see Figure 1(b)). This weighting tends to emphasize visual words that are stable and reliable, while also suppressing spurious visual words.

Consider a visual vocabulary,\( W \), consisting of \( N \) visual words. In addition, consider a set of visual word priorities, \( C = \{ c_1, c_2, \ldots, c_N \} \), where each \( c_i \) represents a score that is proportional to the validity of the visual word. In the absence of any prior information, we assign \( c_i = 0, \forall i \). We then carry out pairwise geometric verification, using a robust estimator to identify a set of inliers. We then update the priority of the inlier visual words based on the results of this process. In the simplest possible scheme, for each feature match that was found to be an inlier, we update a count \( c_i \) for the corresponding visual word. Intuitively, over time, we expect that these counts will help identify visual words that are frequently matched as inliers, as well as words that repeatedly fail the geometric consistency check. This weighting of visual words can then be easily incorporated into a RANSAC framework that biases the sampling in favour of the more reliable words. In our experiments, this resulted in a significant computational speedup.

2.2. Identifying landmark images

As a second contribution, we show that it is possible to learn additional useful information capturing higher-level information about the dataset. For instance, once we have obtained a sufficiently large set of successfully verified image pairs, we hypothesize that this set captures useful information about the global appearance of various landmarks present in the dataset. This information can then be used to train a classifier that distinguishes between landmark and non-landmark images (see Figure 2). Once trained, we can use this classifier in the registration stage. In other words, we first run the classifier on each image before verification. If the classifier has a positive response we continue with geometric verification, but if the response is negative we reject the image immediately, thus significantly reducing the overall compute time.

3. Conclusion

In summary, this work presents techniques for taking advantage of the information generated during geometric verification, to improve the overall efficiency of the process. Our approach thus integrates online knowledge extraction seamlessly into structure-from-motion systems, and is particularly relevant for large-scale image collections. Our results demonstrate both improved efficiency, as well as higher image registration performance, potentially yielding more complete 3D models for these large-scale datasets.

References