Steps Towards A Large-Format Document Hand-Held Scanner

Adrian Ilie

Computer Science Department
University of North Carolina at Chapel Hill [*]

Abstract

There is no easy way to scan large-format documents such as posters, maps, whiteboards, etc. This report describes an implementation of some of the mosaicing techniques described by [Brown and Lowe 2003] and [Capel and Zisserman 1998] for automatically constructing panoramas in the hope of applying them to reconstructing large-format documents from a series of partial snapshots taken independently or with a video camera.

Keywords: video mosaics, panoramas, scale-invariant features

1 Introduction

Methods for scanning large-format documents include scanning them piece-by-piece with a flatbed scanner or scanning them with a hand-held scanner, then stitching the results. The first approach gives excellent resolution, but not all documents can be scanned piece-by-piece. Using the second approach, wall-mounted documents may also be scanned, but the method requires close contact with the document (the hand-held scanner has to come in contact with the document's surface), and is often impractical even for moderately-sized documents.

Moreover, such large-format documents often need not be scanned at the highest resolution possible, but only at enough resolution to allow reading or performing automatic Optical Character Recognition on them.

The techniques used in panoramas readily apply themselves to this problem. Given several snapshots of a large-format document, taken with a regular camera or a video camera, an image of the poster can be reconstructed, even at a higher resolution than any of the input images. While the resolution is not comparable to the one obtained from using a scanner, the ease of use of this method makes it suitable for cases when only a low resolution (such as enough to allow readability) is sufficient.

Stitching several images together, also called *mosaicing*, involves registering the images. In the case of scanners, it usually involves rotations in the same plane. In general, stitching is possible for images that are related to each other by a global mapping such as a planar homography. The next sections describe the registration process. The steps of the process are: feature matching, image matching, image warping and blending.

2 Feature Matching

Most methods for automatic image matching are based on matching features between the images. Correspondences are established between points, lines, or other geometrical entities.

The approach used in [Capel and Zisserman 1998] is extracting Harris corners [Harris and Stephens 1988] and using normalized cross-correlation of local intensity values to match them. However, Harris corners are not invariant to scaling, and cross-correlation is not invariant to rotation.

For this application I used Scale Invariant Feature Transform (SIFT) features [Lowe 1999] instead of Harris corners. SIFT features are designed as geometrically invariant under similarity transforms and also invariant under affine changes in image intensity.

I first tried to implement the approach from [Lowe 2003] myself, and got as far as computing the key locations and magnitudes, but ran into difficulties when computing the key orientations and descriptors. I ended up using the implementation made available online by the authors, which takes a gray-scale image as input and outputs a text file containing the location, scale, orientation and key descriptors of the SIFT features.

Figure 1 shows a comparison between the key locations I computed, and the ones computed using the online implementation. The locations are sometimes different because of the selection criterion: I selected keys based on thresholding the Difference-of-Gaussian pyramids, while the online method also uses some other parameters and processes that were not available for comparison.

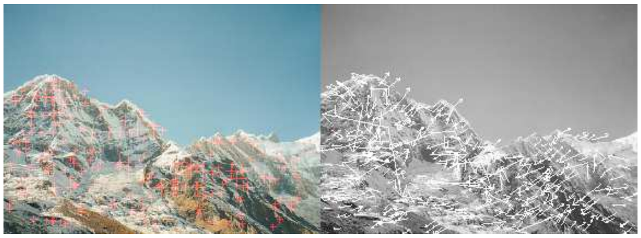

Figure 1: A comparison between two methods of finding keys: mine (left) and the one online (right).

Each image typically contains several hundred features. Since a feature may appear in multiple images, features from different images must be matched against each other. The key descriptors are vectors of 128 numbers. This allows an efficient implementation of the matching that reduces the computation time from $O(n^2)$ to $O(n \log n)$ by constructing a k-d tree with the descriptor values.

The authors of [Beis and Lowe 1997] have shown that given a key, an approximate matching descriptor can be found faster using the k-d tree, and the approximation is good enough for practical applications.

The matching criterion that I used was that the distance from the key to the closest descriptor in the k-d tree is less than a fraction (usually 0.6) of the distance from the key to the second closest descriptor in the k-d tree i.e., the key is closer to the match than to any other descriptors.

The descriptors are indeed very robust: if a match is found between a pair of SIFT features, it is almost certainly not a spurious match. Figure 2 shows a few of the computed correspondences between two images.

3 Image Matching

The next step is matching the images that contain a common subset of features. Feature matching helped identify the images that have

[*] email:adyilie@cs.unc.edu

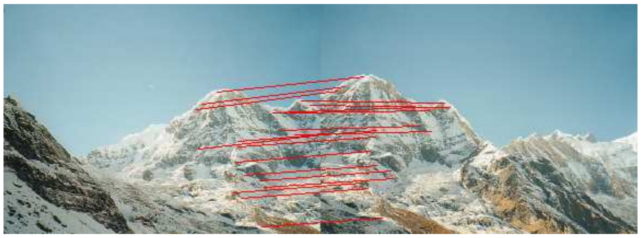

Figure 2: Correspondences between two images.

a large number of matches between them. The authors of [Brown and Lowe 2003] derive and use a probabilistic model, whereby two images match if $N_{inliers} > 5.9 + 0.22\ N_{outliers}$.

They use this criterion to test if pairs of images with many matches have enough matches that are geometrically consistent with a RANSAC Homography. I used MLESAC [Torr and Zisserman 2000] instead of RANSAC to estimate the homographies between such pairs of images. Figure 3 shows an example of a homography used to warp an image so that it overlaps with another image.

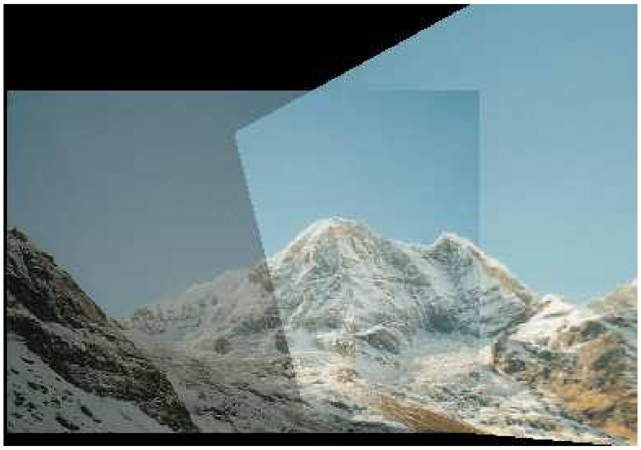

Figure 3: Image warping using a homography (un-warped image on the left is shown semi-transparently).

This step is useful because it helps eliminate false matches, so that panoramas can be automatically extracted from arbitrary image sequences.

4 Bundle Adjustment

Note : I have not finished implementing this yet.

Concatenation of pairwise homographies usually cause accumulated errors. To solve this problem, bundle adjustment can be used to solve for all of the camera parameters jointly, given a set of geometrically consistent matches between pairs of images. Images are added to the bundle adjuster one by one, with the best matching image (maximum number of matches) being added at each step. The new image is initialized with the same rotation and focal length as the image to which it best matches. Then the parameters are updated using Levenberg-Marquardt [Kanzow et al. 2002]. The objective function we use is a robustified sum squared projection error. That is, each feature is projected into all the images in which it matches, and the sum of squared image distances is minimized with respect to the camera parameters.

5 Multi-band Blending

Note : I have not finished implementing this yet.

Due to changes in aperture/exposure time, vignetting (intensity decreases towards the edge of the image), parallax effects due unwanted motion of the optical center, and any mis-registration errors due to mis-modelling of the camera, radial distortion etc., pixels from different images that represent the same location in space do not have the same intensity in different images. A simple approach would be to perform a weighted sum of the image intensities along each ray using some weight functions that take into account contributions from all images that overlap an area. However, this can cause blurring of high frequency detail. To prevent this, the authors of [Brown and Lowe 2003] suggest a multi-band blending strategy developed by Burt and Adelson [Burt and Adelson 1983].

The idea behind multi-band blending is to blend low frequencies over a large spatial range, and high frequencies over a short range. This can be performed over multiple frequency bands using a Laplacian Pyramid. According to the authors of [Brown and Lowe 2003], even a simple 2 band scheme is sufficient to yield good results. A low pass image is formed with spatial frequencies of wavelength greater than 2 pixels relative to the rendered image, and a high pass image with spatial frequencies less than 2 pixels. The low frequency information is blended using a linear weighted sum, and the high frequency information is selected from the image with the maximum weight.

6 Discussion and Future Work

There are a number of unanswered questions when applying panorama techniques to posters.

First, I have no sense of what would be the best way to take the snapshots to ensure the best possible result. A possible sequence would be to take an overview snapshot, then zoom in to capture the details.

Second, there is no obvious way to choose the surface on which all images are projected. Detecting the document's corners (assuming it is rectangular) and warping the initial overview snapshot to a rectangle would probably be a good initial step. However, the detail images may not map so well to this warped version.

Third, the level of detail of the final image should be at least as high as to capture the highest level of detail available from the input images, if not more. However, picking a level of detail from the beginning, or refining it during the process is not an obvious choice, as both may work equally well (or equally bad) in different cases.

Fourth, documents with parts that look similar will pose another problem: there is not enough information in a detail image to be able to accurately place it in the result. This is where other knowledge should come into play: it need not be assumed, as in [Brown and Lowe 2003], that the captured images can be in any order. In reality, temporal coherence is the rule, not the exception, and it provides valuable information that should be used, not ignored.

Finally, I haven't even begun to explore the second part of what would help make these techniques suitable for everyday scanning: super-resolution. The authors of [Capel and Zisserman 1998] suggest the use of ML and MAP estimators both for achieving super-resolution from multiple images, and for the improving the estimates of the homographies between the images. I suspect that the sequence of image acquisition that I suggested earlier, combined with a good method of exploiting this knowledge of temporal coherence may help yield better results.

7 Conclusion

In retrospective, a fully-functional large-format document scanner was an overly-ambitious goal for a semester project (maybe a team of a few students would have been appropriate). I ended up implementing parts of the process using code from VXL [Mundy et al. 2003] and double-checking with short MatLab [MathWorks, Inc. 1984–2002] programs. VXL is a great resource, but its lack of documentation makes it difficult to use. I also had to modify the library code in some places to account for degenerate cases that were generating errors.

I have tried the elements I implemented on regular photographs (the figures in this report use two images from [Brown and Lowe 2003]) and on snapshots of posters, with similar results. Surprisingly enough, poster snapshots taken with small angle changes (that is, always having the camera approximately perpendicular to the poster plane) yielded more numerical errors in computing the homographies than snapshots taken from the same point, but at different angles.

The main conclusion of this experiment is that it is feasible to use panorama techniques for scanning large-format documents, but a robust implementation is non-trivial.

References

BEIS, J., AND LOWE, D. 1997. Shape indexing using approximate nearest-neighbor search in highdimensional spaces. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, 1000–1006.

BROWN, M., AND LOWE, D. 2003. Recognizing panoramas. In *Proceedings of ICCV 2003*, 1218–1225.

BURT, P. J., AND ADELSON, E. H. 1983. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics 2*, 4, 217–236.

CAPEL, D., AND ZISSERMAN, A. 1998. Automated mosaicing with super-resolution zoom. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 885–891.

HARRIS, C., AND STEPHENS, M. 1988. A combined corner and edge detector. In *Proc. of the 4th Alvey Vision Conference*, 147–151.

KANZOW, C., YAMASHITA, N., AND FUKUSHIMA, M., 2002. Levenberg-marquardt methods for constrained nonlinear equations with strong local convergence properties, available at http://citeseer.nj.nec.com/596808.html.

LOWE, D. 1999. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, 1150–1157.

LOWE, D., 2003. Distinctive image features from scale-invariant keypoints. Submitted to the International Journal of Computer Vision, available at http://www.cs.ubc.ca/˜ lowe/papers/ijcv03-abs.html.

MATHWORKS, INC., 1984–2002. MatLab, available at http://www.mathworks.com/.

MUNDY, J., WHEELER, F., PERERA, A., FITZGIBBON, A., SCHAFFALITZKY, F., VANROOSE, P., COOTES, T., AND SCOTT, I., 2003. VXL (the vision-something-libraries), available at http://vxl.sourceforge.net/.

TORR, P., AND ZISSERMAN, A. 2000. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding 78*, 138–156.