

Document Capture using Stereo Vision

Adrian Ulges
U. of Kaiserslautern
67663 Kaiserslautern,
Germany
ulges@iupr.org

Christoph H. Lampert
DFKI GmbH
67608 Kaiserslautern,
Germany
lampert@iupr.org

Thomas Breuel
DFKI GmbH¹
67608 Kaiserslautern,
Germany
tmb@iupr.org

ABSTRACT

Capturing images of documents using handheld digital cameras has a variety of applications in academia, research, knowledge management, retail, and office settings. The ultimate goal of such systems is to achieve image quality comparable to that currently achieved with flatbed scanners even for curved, warped, or curled pages. This can be achieved by high-accuracy 3D modeling of the page surface, followed by a “flattening” of the surface. A number of previous systems have either assumed only perspective distortions, or used techniques like structured lighting, shading, or side-imaging for obtaining 3D shape. This paper describes a system for handheld camera-based document capture using general purpose stereo vision methods followed by a new document dewarping technique. Examples of shape modeling and dewarping of book images is shown.

Categories and Subject Descriptors

I.4.1 [Image Processing and Computer Vision]: Digitization and Image Capture

General Terms

Algorithms

Keywords

Stereo Vision, Camera-Based Document Capture, Dewarping

1. INTRODUCTION

Capturing high quality images of documents currently requires large, stationary devices like flatbed, book scanners, or overhead scanners. The long-term goal of our work is to enable ubiquitous and simple document capture: paper documents should be captured automatically and easily using a

¹German Research Center for Artificial Intelligence, IUPR Research Group, www.iupr.org

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DocEng'04, October 28–30, 2004, Milwaukee, Wisconsin, USA.
Copyright 2004 ACM 1-58113-938-1/04/0010 ...\$5.00.

simple gesture, for example, by holding the document near a document camera, or even fully automatic, high-quality capture of all documents that users interact with on their desks. Other potential applications of our work are improved overhead scanners, in particular for fragile and/or valuable documents and books.

When scanning documents with a scanner, both geometry and illumination of the scanning process can be tightly controlled: the document is pressed against a glass surface and a scanning assembly incorporating both a lamp and a line scanner is moved across the page at a constant distance and speed. When capturing documents with a handheld digital camera, both geometry and light reflected from the book's surface vary greatly across a page. For example, when taking digital photographs of an open book, the pages of the book are usually curled and the pixels near the book spine become substantially darker. Intensity variations can be normalized using a number of well-known adaptive thresholding techniques. This paper addresses the problem of modeling the 3D shape of a page, followed by a flattening or dewarping of the page image using that 3D shape information.

There exists considerable prior work in the area of camera-based document capture. An important special case is the capture of planar page images with a camera whose optical axis is not perpendicular to the page surface. This results in perspective distortions, which can be determined from observing size variations or the perspective convergence of text lines [4, 5]; once the parameters of the perspective distortion have been determined, the dewarping step is a simple global operation.

When the page surface is itself curved, then both determining a 3D model of the page surface and dewarping it are more complex. Some prior work exists. For example, using shape from shading techniques, Zhang [7] model the 3D shape of books pressed to a flatbed scanner and use that 3D information to flatten the scan. However, such controlled illumination techniques would seem difficult to apply to handheld camera settings. Several authors have used structured light in assisting in the 3D reconstruction of page shape [1, 2, 6]; while less constrained, such technique are likely still impractical for simple handheld capturing applications, and distracting even for stationary applications in an office setting.

This paper describes a system using general purpose 3D computer vision techniques for determining the 3D shape of an arbitrary page surface from two (or more) images, followed by a method for flattening such surfaces. By not requiring prior camera calibration or structured light, the

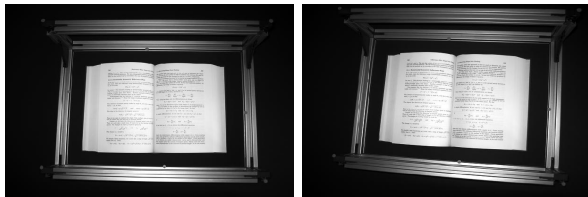


Figure 1: A typical pair of images captured with a handheld consumer digital camera.

system can be applied in a variety of settings. For example, a user can casually capture two or more images using a digital camera, PDA, or cell phone (of sufficient resolution) and process them later. An office setup can consist of two cameras mounted above a user’s desk, whose images are processed continuously by the user’s desktop PC. And we can imagine mass-producing a handheld document camera, comparable in size to a current handheld digital camera, consisting of a stereo capture setup (e.g., two CMOS image sensors) and the software described in this paper.

2. METHODS

Hardware. Our goal in this work has been to avoid the use of specialized hardware or prior camera calibration. For image capture, we have been using a Canon Powershot S50, a consumer digital camera with 5 Megapixel resolution. Without known reference points in the image, stereo algorithms applied to two images yield projective coordinates in the general case. These can be transformed into Euclidean coordinates either using reference points with known distances, a third image, or prior knowledge about documents (e.g., parallelism of text lines). Here, we just cover the simplest case, in which some known 3D reference points are located in the image (this is still suitable for overhead desktop scanning). Typical images obtained look as in Figure 1.

Epipolar Geometry. Using the reference points (or page-intrinsic information), we first compute an approximation of the image pair’s epipolar geometry using the linear 8-point algorithm and normalized coordinates[3]. Using this epipolar geometry, we warp the raw images into two “rectified” views, views with a known, simple epipolar geometry. In these rectified views, each epipolar line corresponds (approximately) to a line of $y = \text{const}$. While the image warping step is not strictly speaking necessary, it substantially reduces the computational cost of the subsequent correlation-based stereo matching algorithm and improves the quality of the resulting disparities.

Feature Matching. The following search for correspondences between the two images is the most critical part in terms of runtime. To be as close as possible to a real-time system, we use fast block matching with SIMD optimized assembler routines which are part of the MPEG-4 video codec Xvid¹. For the actual search, we split the left rectified image into square blocks of typically 16×16 or 32×32 pixels. To avoid unreliable matches we exclude all blocks that do not contain strong corners, as determined by a standard corner

¹Source code available under GNU General Public License from www.xvid.org.



Figure 2: Feature blocks and displacement. Blocks with little or no structure in their texture are left out, to reduce the number of unreliable matches.

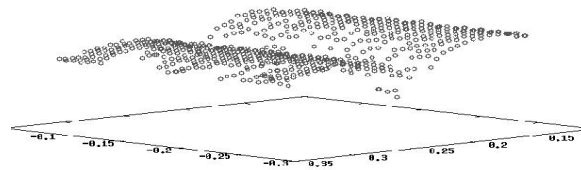


Figure 3: Reconstructed 3D points from 750 surface matches. A gridlike structure automatically appears due to the choice of feature points on a regular grid.

detector. For each block, we search along its center’s epipolar line in the second image for the position with highest correlation. (Figure 2). To reduce the number of outliers, we then repeat the whole matching procedure with the right image, bringing it into correspondence with the left. Of the resulting two sets of matches, we only keep those that are symmetrical between both runs.

3D Reconstruction. The previous step yields a set of usually between 500 and 2000 displacement (disparity) vectors. Our system can either use these disparity vectors directly for computing depth-from-stereo, or permits a second step for refining the epipolar geometry, followed by depth computation. A resulting 3D book surface is shown in Figure 3.

Fitting and Flattening. Because of its physical properties, the class of 3D surfaces usually formed by paper is in a class called *applicable surfaces* (e.g., [6]), surfaces with zero Gaussian curvature. When such surfaces are approximated by a regular mesh, the resulting mesh can be approximately flattened without changing the lengths of the mesh lines. To obtain such a mesh, we use a two-step process. The first is to reconstruct a surface using a moving average, robust least square fitting procedure. The resulting surface is smooth but not necessarily applicable. We then construct a least square fit of a plane through the surface. Within this plane, we construct a regular mesh and project it onto the actual surface. In general, this will result in an irregular mesh on the depth surface. Finally, we perform stochastic gradient descent, minimizing the total square deviation of edge lengths from the regular mesh’s correct edge length, to obtain a regular mesh approximating the book’s 3D surface.

Image Dewarping. The triangles in the regular mesh on the book’s surface have a one-to-one correspondence to any regular planar mesh with the same number and connectivity of nodes (the regular mesh that we constructed in the least

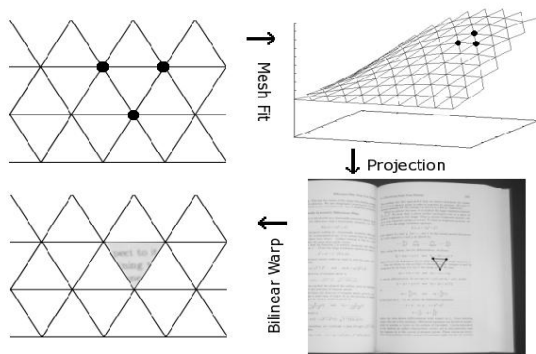


Figure 4: Image Dewarping. For a triangle within the planar mesh (top left), the 3D coordinates of its edges in 3D space (top right) are projected to the original image. The resulting triangle defines the texture patch (bottom right) which is warped bilinearly to its target position (bottom left).

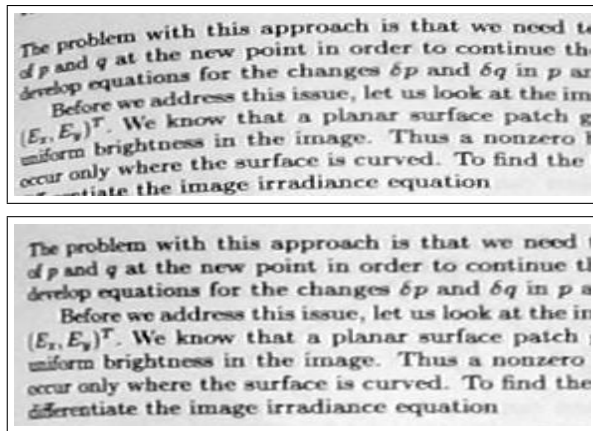


Figure 5: Image region before and after dewarping. The curl has been removed almost completely.

square fitting plane is an example of such a mesh). In order to perform dewarping, we consider each triangle in these regular meshes. We take the coordinates of the 3D triangle, project it into one of the two images (e.g., the left). We use the image pixels contained within the projection of that triangle and warp it onto the corresponding triangle in the planar regular mesh. After we have completed this operation for all triangles in the mesh, we obtain a dewarped representation of the original book surface. This procedure is illustrated in Figure 4.

An enlarged portion of a dewarped image from applying this procedure to the image pair in Figure 1 are shown in Figure 5. Similar results are obtained with other curved documents, both books and curled pages.

3. DISCUSSION

This paper has described a collection of techniques and demonstrated a complete system for capture and dewarping of document images using techniques from 3D computer vision. The method is potentially applicable to a wide range of document types, settings, and camera geometries.

The use of a stereo algorithm means that we can dispense with structured lighting or other intrusive techniques. In addition, by using a stereo algorithm that works with uncalibrated cameras, we avoid the need for careful geometric adjustment. That means that, in stationary setups, two cameras can be mounted above a work surface, covering approximately the same field of view, and the system can infer all its internal parameters from a one time presentation of a test pattern. This eliminates the hardware adjustments associated with traditional stereo camera setups.

Furthermore, the use of such stereo algorithms permits the use of handheld digital cameras for document capture. Using our current system, this still requires the presence of known points in both images, say, by identifying a few feature points around the user’s desk. But that requirement can be replaced using knowledge of geometric features commonly found in document images, which will permit simple, high-quality document capture anywhere, using no more than a regular digital camera.

Image quality can also be improved in some areas. One of the most important areas is that, close to the spine of books, there often remains a small amount of geometric distortion even in the dewarped image. The cause of this is a bias in the initial surface interpolation procedure: the line where the two facing pages of a book meet represents a significant discontinuity in the derivative of the depth surface, but the simple smoothing procedure we use currently assumes a smooth first derivative. This results in an underestimate of the depth near the book spine in the interpolated surface, and as a result, the dewarping at that location is incomplete. Known surface interpolation techniques can be used to improve the depth surface interpolation and will likely eliminate this effect.

A careful performance evaluation should, of course, follow such improvements, and eventually, a standardized database of document images captured with handheld cameras may permit a comparison with other approaches.

In the medium term, we plan to incorporate the software into actual devices: desktop scanning and prototypes of self-contained portable handheld document cameras are the most immediately useful applications.

4. REFERENCES

- [1] M. S. Brown and W. B. Seales. Document restoration using 3d shape: A general deskewing algorithm for arbitrarily warped documents. In *International Conference on Computer Vision (ICCV01)*, volume 2, pages 367–374, July 2001.
- [2] P. Cubaud, J.-F. Haas, and A. Topol. Numérisation 3d de documents par photogrammétrie. In *Actes du Huitième – Colloque International Francophone sur l’Ecrit et le Document*, pages 291–296, June 2004.
- [3] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proc. IEEE Conf. on Computer vision and Pattern Recognition*, pages 761–764, 1992.
- [4] W. Newman, C. Dance, A. Taylor, S. Taylor, M. Taylor, and T. Aldhous. Camworks: A video-based tool for efficient capture from paper source documents. In *IEEE International Conference on Multimedia Computing and Systems - Volume 2*, pages 647–653, June 1999.
- [5] M. Pilu. Deskewing perspective distorted documents: An approach based on perceptual organization. *HP White Paper*, May 2001.
- [6] M. Pilu. Undoing page curl distortion using applicable surfaces. In *Computer Vision and Pattern Recognition Conference*, pages 67–72, December 2001.
- [7] Z. Zhang. Restoration of curved document images through 3d shape modeling. In *International Conference on Computer Vision and Pattern Recognition (CVPR2004)*, June 2004.