21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Robust Stereo Correspondence for Documents by Matching Connected Components of Text-Lines with Dynamic Programming

Martin Krämer, Muhammad Zeshan Afzal, Syed Saqib Bukhari
Technical University Kaiserslautern
{kraemer,afzal,bukhari}@iupr.com

Faisal Shafait
German Research Center for Artificial Intelligence (DFKI)
faisal.shafait@dfki.de

Thomas M. Breuel
TU Kaiserslautern
tmb@iupr.com

Abstract

In this paper we present a novel method for robust stereo matching on document image pairs. The matching itself is performed using an affine-invariant similarity measurement to compensate for perspective distortions, where affine invariance is achieved by normalization using second-order statistics, to finally allow a simple pixel-wise comparison. To handle the inherent high self-similarity of the page content we apply a dynamic programming approach on text-line pairs. We quantitatively show that the proposed method performs better in comparison to standard approaches using SURF – whether with or without incorporating text-line information.

1. Introduction

The research community is increasingly interested in document archival systems based on camera-captured images, e.g. [11], [10]. It both allows for faster and easier capturing compared to a conventional flat-bed scanner and does not require mechanical manipulation of the documents in question. The difficulty of the procedure is that the captured data needs to be visually transformed to appear as if it would have been captured by a flat-bed scanner, i.e. dewarping, to allow both for easier reading as well as further software processing like OCR.

For good performance of the dewarping approach it is beneficial to determine the three-dimensional shape of the captured object. Due to the high self-similarity of document images it is not trivial to achieve an acceptable correctness and robustness of the algorithm. The presented approach is based on the general idea of Ohta [8], who applies dynamic programming on scan-lines of rectified images for general purpose 3D-reconstruction of stereo images. Instead of working with the rectified scan-lines we use the document's text-lines to apply dynamic programming on and connected components serve as features instead of a pixel (or window thereof). Regarding text-line extraction one of the currently best performing methods have been published by Bukhari et al. [4]. In this paper we apply a convolution of isotropic Gaussian filters with a set of line filters [5], which produces good results. To the best of our knowledge the only publication, which tries to improve stereo correspondence of document images using text-line information, has been published by the authors [1]. There we only restricted the matching process using the predetermined assignments of detected features to text-lines. The general idea of exploiting domain-specific knowledge is commonly applied to obtain higher quality 3D-reconstructions. Beeler et al. [3] incorporated constraints based on human face geometry, e.g. smoothness, to facilitate high-quality reconstructions using only a consumer-grade stereo camera.

There are no relevant publications regarding matching of connected components for stereo correspondence on document images, that directly try to solve the problem of perspective distortions, but the work by Keysers et al. [7] follows a conceptually close idea. They are modifying similarity measurements in such a way, that certain expected deformations in the input image compared to the prototypes are cost-free. Although their work is geared towards recognition of hand-written characters, the underlying idea seems equally applicable for comparing connected components from stereo images, which have undergone perspective transformations.

978-4-9906441-0-9

2. Preprocessing

Necessary preprocessing steps of the presented approach span splitting of the book pages, binarization, removal of "bad" connected components, background cleaning, text-line extraction/labeling and establishment of text-line correspondences. The book pages are currently split in a manual step as we are yet lacking an automatic procedure.

For binarization we use the local adaptive thresholding method by Sauvola [9] to compensate for local intensity differences introduced by the capturing process. All connected components, which are unlikely to represent a cleanly extracted character are removed. Relevant criteria for this step are the area and aspect ratio of their bounding boxes. This clean-up step avoids that large series of merged characters or layout elements, like horizontal lines, are kept, which would interfere in the matching process. Although this step works very well in practice, it still fails to remove all background artifacts in some images, which were cleaned manually.

Text-lines are afterwards detected by a method proposed by Bukhari et al. [5] using a convolution of isotropic Gaussian filters with a set of line filters. This yields a labeling of connected components according to text-line correspondences. For determining text-line correspondences we currently apply a naive procedure, which uses ordering and area information of text-lines to alleviate under- or over-segmentation [1].

Although this approach works quite well in practice, it is not error-free and we fixed the small number of remaining errors manually to focus on the performance of stereo correspondence. A more robust algorithm for finding line correspondences would be necessary for practical application.

3. Stereo Correspondence

3.1 Matching

The general problem with camera-captured images are the perspective distortions introduced by the capturing process. For the case of stereo correspondence on document images this prohibits simply extracting connected components (or characters) and trying to match them directly, because they are generally captured from a different viewpoint.

Suppose we have a connected component C_l from the left image and a connected component C_r from the right image, which are always converted to a fixed size before further processing. For sake of simplicity we may omit the $_l$ and $_r$ indices, when the equations are applicable for both. $C = (i_{xy})$ denotes a connected component and contains the image intensities i_{xy} at x, y coordinates. Now we first determine correlations s_{xx}, s_{xy}, s_{yy} between image intensity and direction to find the character orientation in both images:

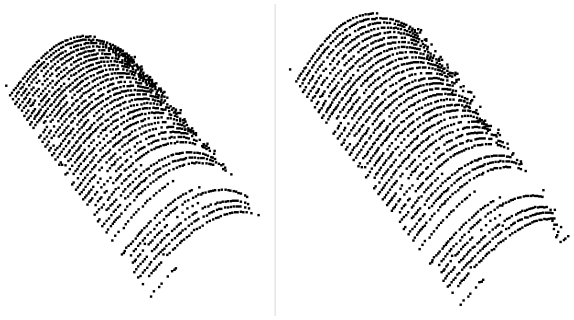

Figure 1: 3D reconstruction of book page. Left side shows results after application of RANSAC. Right side shows original output.

$$
\begin{aligned}
s_x &= \textstyle\sum_{x,y} i_{xy} \times x \qquad s_y = \sum_{x,y} i_{xy} \times y \\
m_x &= s_x / \textstyle\sum_{x,y} i_{xy} \qquad m_y = s_y / \sum_{x,y} i_{xy} \\
s_{xx} &= \textstyle\sum_{x,y} i_{xy} \times (x - m_x)^2 \\
s_{yy} &= \textstyle\sum_{x,y} i_{xy} \times (y - m_y)^2 \\
s_{xy} &= \textstyle\sum_{x,y} i_{xy} \times (x - m_x) \times (y - m_y)
\end{aligned}
$$

After this we are now able to compute the affine transformation R, which transforms C_r into $C_{r-norm} = R(C_r)$ such that perspective distortions introduced by different capturing viewpoints are compensated for. Therefore we decompose the matrices

$$M_l = \begin{pmatrix} s_{l,xx} & s_{l,xy} \\ s_{l,xy} & s_{l,yy} \end{pmatrix} \text{ and } M_r.$$

Let $V = (v_1 v_2 v_3)$ denote the unit length eigenvectors of M in column-order and e_1, e_2, e_3 be the corresponding eigenvalues. Then we decompose M into

$$M^d = V \cdot diag(e_1^{0.5}, e_2^{0.5}, e_3^{0.5}) \cdot V^T.$$

Finally the affine transformation R can now be determined by $R = M_l^d \cdot M_r^{d^{-1}}$.

To negate interpolation artifacts, which may occur due to the affine transformation, we apply greyscale dilations with differently sized rectangular structuring elements, i.e. we used the original image as well as $n \times n$ structuring elements ($1 \leq n \leq 3$), on either C_l or C_{r-norm}, depending on which one exhibits a smaller sum of intensity values, and remember the sum of absolute differences for each comparison with its counterpart. The minimal error resulting from this procedure gives the final matching score c_n.

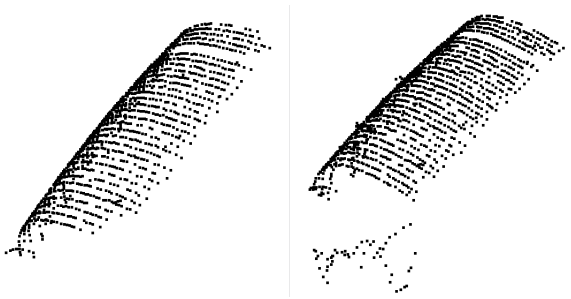

Figure 2: 3D reconstruction of book page with matching errors. Left side shows results after application of RANSAC. Right side shows original output.

3.2 Dynamic Programming

To handle the problem with ambiguous matching we pair up corresponding text-lines of the stereo image pair and apply a dynamic programming approach on each pair. The underlying graph is hereby built as following: first we insert dummy start and end nodes to allow application of Dijkstra's algorithm.
Each non-dummy node (i_l, i_r) denotes a match of one connected component from the left text-line, i.e. indexed by i_l, and one from the right text-line, i.e. indexed by i_r. Valid non-dummy start nodes are all (i_l, i_r) such that $0 \leq i_l < max_{\text{skip}}$ and $0 \leq i_r < max_{\text{skip}}$, where max_{skip} is a constant defined to avoid skipping too many connected components. Similarly valid non-dummy end nodes are defined such that they are no further from the last node than max_{skip} steps. Now valid successors for a given node (i_l, i_r) are given by (j_l, j_r) such that $i_l < j_l \leq i_l + max_{\text{skip}}$ with an analogous constraint on j_r. The full graph is now constructed by adding valid successors to all non-dummy start nodes recursively until only non-dummy end nodes are reached. Furthermore all nodes, where the aspect ratio of matched connected components would differ by more than a defined constant, are removed from the final graph.
Now each edge of the graph gets a cost assignment to reflect match quality and allow computation of optimal matching. The cost is composed of a node cost c_n, which reflects the quality of matching the connected components, a skip cost c_s, which penalizes skipping of connected components and a roughness cost c_r, which penalizes strong divergences of neighboring disparities. Calculation of c_n is explained in the last subsection. For an edge between nodes (i_l, i_r) and (j_l, j_r) we define

$$c_s = c_1 \times ((j_l - i_l - 1) + (j_r - i_r - 1))^2,$$

where c_1 is a user-defined constant.

$$c_r = c_2 \times abs(d_1 - d_2)^2$$

is given by squaring the absolute difference of disparity d_1 of first node and disparity d_2 of second node before multiplying with constant c_2. The disparities are determined by measuring the horizontal difference in pixels of mass centers of matched connected components across the stereo image pair. Finally the edge cost $c = c_n + c_s + c_r$ is simply set to the sum of its composite costs.
By tracing the minimum cost path through the graph we get the matching of the two given text-lines.

4. Experimental Results

We use the same dataset as in our previous paper [1] to allow comparison of results. It is captured using a standard stereo setup and consists of one hundred stereo image pairs (resolution of ten mega-pixels) of a book page. Robustness of three-dimensional reconstruction is, like in [1], measured by observing the percentage of discarded matches by applying RANSAC [6] on the epipolar constraint, i.e. $p_r^T F p_l = 0$, where p_l is a point from the left image, p_r is a point from the right image and F denotes the fundamental matrix.
For Sauvola binarization of image pairs we set the threshold $k = 0.3$ and window size to 60. Criteria for "good" connected components were defined as following: bounding box area between 50 and 8000 pixels to remove noise and large background artifacts and aspect ratio between 0.2 and 2.5 to get rid of layout elements and large merged character groups. All connected components were converted to a size of 64×64 before matching. We found these settings to work best by empirical experiments.
For the dynamic programming approach we set the parameters as following: skip at most three connected components on every graph transition; do not match connected components, if the ratio of their aspect ratios differs by more than 30%; skip cost coefficient $c_1 = 10000$; roughness cost coefficient $c_2 = 1000$. These parameters were again found to perform best by empirical experiments.
An illustration of a well-working example can be seen in Figure 1. No significant errors occur across the page surface, but minor deviations at the book ridge are present. This happens due to slight differences in connected components at the highly distorted areas across a stereo image, which are still similar enough to allow matching. In Figure 2 an example is given, where matching does not work properly for two text-lines. As their reconstructed 3D points violate the epipolar constraint strongly RANSAC is able to remove the error completely. We verified in practice that both cases, i.e. small deviations at ridge and small missing areas, can be

Method	Good	All	Good %
This Paper	1373.12	1475.38	92.80%
SURF Textlines [1]	1327.78	1946.26	67.75%
SURF [1]	1001.25	3281.63	29.98%

Table 1: Robustness comparison. The number of good and overall matches averaged over the whole dataset are given in columns two and three respectively. A match is regarded as good if it does not violate the epipolar constraint by more than two pixels. The fourth column gives the average inlier vs outlier ratio.

handled in a follow-up step, which fits a general cylinder to the given point cloud with a robust least squares approach.
Results of our experiments are given in Table 1. We can see that the presented approach yields a significantly higher percentage of correct matches in comparison to application of SURF [2] – whether with restraining matching using text-line information or not – while producing slightly more correct matches overall as well.

5. Future Work

There are several possibilities to improve upon the presented approach with regards to practical applicability. To get rid of the existing manual preprocessing steps firstly a more robust method for establishment of text-line correspondences, secondly automatic separation of book content from background and thirdly separation of the two book pages have to be incorporated into the presented approach. Furthermore a way of handling non-text regions would be necessary to deal with cases, where not much text is present, to yield an acceptable coverage of features, which is fundamental to achieving good three-dimensional reconstruction.

6. Conclusion

We presented a novel method for stereo correspondence on document images. The matching is performed with a simple comparison of connected component images after normalizing size and neutralizing effects of affine transformations. By restricting the matching to corresponding text-lines and applying a dynamic programming approach on the text-line pairs we are able to deal with the high self-similarity of the page content. Differences in the connected components after binarizing the image pair are handled by allowing skips in matching, when no good match can be established.
Overall the method yields a superior robustness compared to our previous experiments with SURF, restricted by text-lines [1], and seems to be applicable in practice after incorporating the improvements outlined in the previous section.

References

[1] M. Afzal, M. Krämer, S. Bukhari, F. Shafait, and T. Breuel. Improvements to Uncalibrated Feature-based Stereo Matching for Document Images by using Text-Line Segmentation. In *Proceedings of the 10th IAPR International Workshop on Document Analysis Systems*, 2012.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding*, 110:346–359, 2008.

[3] T. Beeler, B. Bickel, P. Beardsley, R. Sumner, and M. Gross. High-Quality Single-Shot Capture of Facial Geometry. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 29(3):40:1–40:9, 2010.

[4] S. Bukhari, F. Shafait, and T. Breuel. Performance Evaluation of Curled Textline Segmentation Algorithms on CBDAR 2007 Dewarping Contest Dataset. In *Proceedings of the 17th IEEE International Conference on Image Processing*, pages 2161–2164, 2010.

[5] S. Bukhari, F. Shafait, and T. Breuel. Text-Line Extraction using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters. In *Proceedings of the 11th International Conference on Document Analysis and Recognition*, pages 579–583, 2011.

[6] M. Fischler and R. Bolles. Readings in Computer Vision: Issues, Problems, Principles, and Paradigms. chapter Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, pages 726–740. 1987.

[7] D. Keysers, C. Gollan, and H. Ney. Local Context in Non-linear Deformation Models for Handwritten Character Recognition. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 4, pages 511–514, 2004.

[8] Y. Ohta and T. Kanade. Stereo by Intra- and Inter-Scanline Search using Dynamic Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:139–154, 1985.

[9] F. Shafait, D. Keysers, and T. Breuel. Efficient Implementation of Local Adaptive Thresholding Techniques Using Integral Images. In *Proceedings of the 15th Document Recognition and Retrieval Conference*, volume 6815. SPIE, 1 2008.

[10] A. Ulges, C. Lampert, and T. Breuel. Document Capture using Stereo Vision. In *Proceedings of the 2004 ACM Symposium on Document Engineering*, pages 198–200, 2004.

[11] A. Yamashita, A. Kawarago, T. Kaneko, and K. Miura. Shape Reconstruction and Image Restoration for Non-Flat Surfaces of Documents with a Stereo Vision System. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 482–485, 2004.