2012 10th IAPR International Workshop on Document Analysis Systems

# Improvements to Uncalibrated Feature-Based Stereo Matching for Document Images by using Text-Line Segmentation

Muhammad Zeshan Afzal*†, Martin Krämer*†, Syed Saqib Bukhari*, Faisal Shafait‡ and Thomas M. Breuel*
*Image Understanding and Pattern Recognition Group
Technical University of Kaiserslautern, Kaiserslautern, Germany
Email: {afzal,kraemer,bukhari,tmb}@iupr.com
‡Multimedia Analysis and Data Mining Research Group
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
Email: faisal.shafait@dfki.de

***Abstract*—Document images prove to be a difficult case for standard stereo correspondence approaches. One of the major problem is that document images are highly self-similar. Most algorithms try to tackle this problem by incorporating a global optimization scheme, which tends to be computationally expensive. In this paper, we show that incorporation of layout information into the matching paradigm, as a grouping entity for features, leads to better results in terms of robustness, efficiency, and ultimately in a better 3D model of the captured document, that can be used in various document restoration systems. This can be seen as a divide and conquer approach that partitions the search space into portions given by each grouping entity and then solves each of them independently. As a grouping entity text-lines are preferred over individual character blobs because it is easier to establish correspondences. Text-line extraction works reasonably well on stereo image pairs in the presence of perspective distortions. The proposed approach is highly efficient and matches obtained are more reliable. The claims are backed up by showing their practical applicability through experimental evaluations.**

***Keywords*-stereo matching; stereo correspondence; 3D reconstruction; document capturing; feature grouping**

## I. Introduction

Capturing document images using portable cameras is efficient and inexpensive. Irrespective of many advantages, the document images obtained by the camera are degraded by geometric, e.g. perspective distortion, and photometric, e.g. specular highlights, artifacts. Geometric artifacts may, for instance, be introduced by camera imaging models or book pose and photometric artifacts can be caused by the illumination changes by environmental light sources. Robust processing of image documents in the presence of these artifacts is an active area of research.

There are two ways to acquire document geometry. One is to use active sensors, which requires a specialized laboratory setup. Alternatively stereo cameras may be used. Nowadays stereo cameras do not require a special setup, because they are readily available in portable devices and it is very easy for an end user to acquire stereo images with them.

For performing 3D reconstruction using stereo image pairs, the most important part is to establish robust correspondences between features detected in both images. Establishing correspondences is especially problematic for document images because of their highly self-similar content. The general procedure for document images, which has been followed in most approaches, is to detect robust features in both images, e.g. using SIFT or SURF. These features are then matched under constraints established using epipolar geometry, i.e. matching along epipolar lines. After that a suitable model is fitted, depicting book shape information, for further refinement. One such example is the use of NURBS [1] to the obtained point cloud, which should be as close as possible to the ideal book surface. Our proposed approach makes the processing of robust feature matching simple and efficient by exploiting layout information contained in document images without any model fitting.

Among the various available layout elements, the proposed algorithm takes segmented lines as input and features are matched over these text-lines. The feature grouping is highly beneficial for robust matching because potential outliers located on a different text-line are removed from the list of potential match candidates, while keeping good candidates.

The rest of the paper is organized as follows: in section II a brief overview of related research is given. Section III discusses the preprocessing steps. Section IV describes feature extraction and matching, which is used for the experiments discussed in section V. Section VI concludes the paper with some insights on future work.

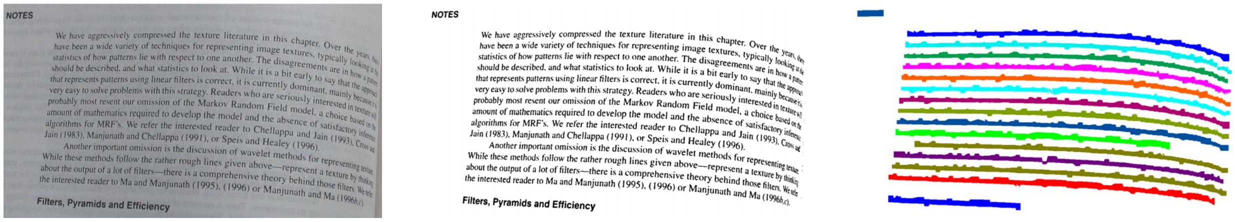
## II. Related Work

As camera captured documents are distorted, it is required to use features that are robust under affine transformations. SIFT [2] and SURF [3] are both well-known and widely used feature descriptors that suit these requirements.

Another critical step for the presented approach is robust extraction of curled text-lines from document images. Techniques that are generally used for document images

† Both authors contributed equally to this work.

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.44



NOTES

We have aggressively compressed the texture literature in this chapter. Over the years, ther
have been a wide variety of techniques for representing image textures, typically looking at the
statistics of how patterns lie with respect to one another. The disagreements are in how a patter
should be described, and what statistics to look at. While it is a bit early to say that the approach
that represents patterns using linear filters is correct, it is currently dominant, mainly because it is
very easy to solve problems with this strategy. Readers who are seriously interested in texture will
probably most resent our omission of the Markov Random Field model, a choice based on the
amount of mathematics required to develop the model and the absence of satisfactory inference
algorithms for MRF's. We refer the interested reader to Chellappa and Jain (1993), Cross and
Jain (1983), Manjunath and Chellappa (1991), or Speis and Healey (1996).
Another important omission is the discussion of wavelet methods for representing texture
While these methods follow the rather rough lines given above—represent a texture by thinking
about the output of a lot of filters—there is a comprehensive theory behind those filters. We refer
the interested reader to Ma and Manjunath (1995), (1996) or Manjunath and Ma (1996b,c).

Filters, Pyramids and Efficiency

NOTES

We have aggressively compressed the texture literature in this chapter. Over the years, ther
have been a wide variety of techniques for representing image textures, typically looking at the
statistics of how patterns lie with respect to one another. The disagreements are in how a patter
should be described, and what statistics to look at. While it is a bit early to say that the approach
that represents patterns using linear filters is correct, it is currently dominant, mainly because it is
very easy to solve problems with this strategy. Readers who are seriously interested in texture will
probably most resent our omission of the Markov Random Field model, a choice based on the
amount of mathematics required to develop the model and the absence of satisfactory inference
algorithms for MRF's. We refer the interested reader to Chellappa and Jain (1993), Cross and
Jain (1983), Manjunath and Chellappa (1991), or Speis and Healey (1996).
Another important omission is the discussion of wavelet methods for representing texture
While these methods follow the rather rough lines given above—represent a texture by thinking
about the output of a lot of filters—there is a comprehensive theory behind those filters. We refer
the interested reader to Ma and Manjunath (1995), (1996) or Manjunath and Ma (1996b,c).

Filters, Pyramids and Efficiency

(a) Original image of book page

(b) Binarized image of the book

(c) Illustration of text-line segmentation

Figure 1. A small portion from one of the sample image has been shown in original, binarized and text-line segmented forms to illustrate preprocessing step

captured by a flat-bed scanner do not perform very well when applied on camera-captured images instead [4]. The reason for this is that text-lines are significantly curled due to the book surface and skewed due to the inherent perspective distortions, imposed by the camera capturing process, which makes robust extraction a challenging task. The work-flow in this paper uses the method proposed by Bukhari et al. [5] for curled text-line detection.

Although techniques for search space reduction in context of feature matching have already been examined in literature since quite some time [6], [7], [8], they differ significantly from our approach. Venkateswar and Chellappa [6] build a hierarchy of image features, i.e. lines, vertices's, edges and surfaces, and use the obtained relationships for constraining feature matching. Horaud and Skordas [7] extract linear edge segments and restrict matching by taking the relationship to neighboring features into account. Another method for search-space reduction is presented by Moallem and Faez [8], which does not impose hierarchical constraints, but rather tries to reduce the set of matching candidates by limiting the allowed disparity range.

In comparison, we are using affine invariant and robust features and our algorithm is specialized for the application on document images. We can exploit accurate higher level layout features for grouping, instead of relying on rather unstable lower-level features. Additionally, we do not require a global optimization scheme, based on graph matching or other approaches, and are able to execute a rather naive approach after subdividing the search space appropriately

Our proposed method can be used in a variety of document restoration systems e.g [9], [10], [11], which require a 3D model of the captured book for removing perspective distortions, to finally produce output similar to a flat-bed scanner. These systems generally rely on sparse feature matching to obtain a 3D point cloud and try to fit a parametrized model, which exploits known geometric properties of book shapes, e.g. smoothness, and compensates for outliers. Improving the initial 3D point cloud will allow such systems to operate on more robust input data and estimate the real book shape more accurately.

## III. Dataset Acquisition and Preprocessing

The data-set for our empirical evaluation has been captured using a standard stereo setup with two cameras, pointing at the same document, from different views in an ortho-parallel setup. It consists of 100 stereo image pairs taken from different books. Images are captured in natural lighting conditions. Images mostly consist of text-lines except some special formatting in header and footer for some images. Preparation of the dataset involves manual steps for removal of background and ridge detection for separating the individual book pages. In a practical setup, document cleaning algorithm e.g [12] could be applied instead.

The preprocessing step includes binarization, text-line detection, application oriented text-line labeling and finding line correspondences. It starts with the binarization of the document images as seen in Figure 1b. Here we use a local adaptive thresholding method, i.e. Sauvola [13] but with the efficient implementation proposed in [14], for binarization. The value for the threshold parameter is set to 0.3 and the window size is set to 70 for all of the images in the dataset for binarization. Text-lines are detected by applying ridge-based curled text-line detection method as proposed by Bukhari et. al. [5]. The text-line extraction algorithm outputs detected text-lines in a color coded format [15], in which each connected component of a text-line is assigned the index of the text-line as its color.

Although, we could only use the features lying on connected components, in context of our application, the features lying between spaces of characters are also equally important, as they use information about their context to be robust. In the next step the algorithm is simply connecting the extremes of bounding boxes of blobs corresponding to the same line. The result of this step can be seen in Figure 1c. For establishing correspondences of the text-lines between stereo images, a naive procedure is applied, which uses ordering and area information of text-lines to alleviate correspondence problems arising as a result of under or over segmentation of the text-lines. In practice text-line segmentation works well, but this approach could be enhanced by using a more robust line correspondence

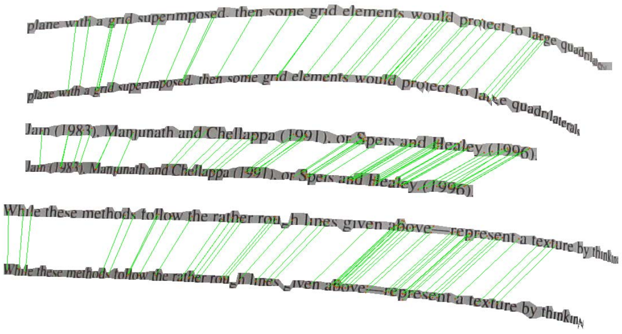

plane with a grid superimposed, then some grid elements would project to large quadrilateral
Jain (1983), Manjunath and Chellappa (1991), or Speis and Healey (1996)
While these methods follow the rather rough lines given above—represent a texture by thinking

Figure 2. Illustration of feature matching

procedure.

## IV. Stereo Correspondence

Stereo correspondence is performed by matching both stereo images against each other. Assuming we have a stereo image pair $I_1, I_2$, where each image has the same width $w$ and height $h$, a stereo correspondence approach can produce a disparity map regarding one or both of the stereo images. The disparity map contains at each position the horizontal offset between matched points and can directly be used for recovering real depth due to their inversely proportional relationship.

We call $d_{ij}(I_1, I_2)$ the disparity map of image $I_1$ in relation to image $I_2$, if the following equation holds:

$$d_{ij} = i - \text{hor}(\text{match}(I_1(p), I_2)) \forall i, j : 0 \leq i < w, 0 \leq j < h$$

hor returns the x-coordinate of a point, i.e. given a point $p = (x, y)$, we define $\text{hor}(p) = x$. match returns the coordinates of the best matching point $p^{'}$ in image $I_2$ given a point $p$ in image $I_1$ by minimizing a cost function $c$:

$$\text{match}(I_1(p), I_2) = \arg \min_{p^{'} \in I_2} c(I_1(p), I_2(p^{'}))$$

Analogously we can define the disparity map of image $I_2$ in relation to image $I_1$ by switching the terms appropriately.

In the following part we will have a closer look at our feature-based approach and explain our strategy for improving robustness and efficiency.

### A. Feature Detection

The first step in a feature-based stereo approach consists of finding and describing distinctive points of interest in both images, which can be matched to recover the disparities.

Performing feature detection on stereo images $I_1$ and $I_2$ results in a set of features per image: $F_1 = \{f_1, ..., f_n\}$ for image $I_1$ and $F_2 = \{g_1, ..., g_m\}$ for image $I_2$. Each detected feature $f = (p, d)$ consists of its spatial position $p = (x, y)$ in the image and a descriptor $d = \{d_1, ..., d_p\}$, where value of $p$ and definition of $d$ depend on the used feature descriptor.

### B. Feature Matching

The task of feature matching step is to establish correspondences between the corresponding feature sets $F_1$ and $F_2$ belonging to left and right image respectively. A suitable method for establishing correspondences is to determine the nearest neighbor of a given feature descriptor $f$ from image $I_1$ among the set of feature descriptors $F_2$ for image $I_2$ using nearest neighbor search techniques e.g [16], [17]. This generally requires definition of the cost function which in our case is Euclidean distance between the feature descriptors of the given points:

$$c(I_1(p), I_2(p^{'})) = ||d_{I_1(p)} - d_{I_2(p')}||_2$$

The matching function $\text{match}(I_1(p), I_2)$ then just returns the spatial position $p^{'}$ of the nearest neighbor $d_{I_2(p')}$ to $d_{I_1(p)}$ in feature space from the set of features in $F_2$:

$$\text{match}(I_1(p), I_2) = \arg \min_{p^{'} \in I_2} ||d_{I_1(p)} - d_{I_2(p')}||_2$$

An illustration can be seen in Figure 2.

As a general matching constraint we ensure that the two best matches $g_{1st}, g_{2nd}$ of a feature descriptor $f_i$ at least have a minimum distance $d_{min}$ between each other. Additionally we enforce that the best match does not exceed a certain distance threshold $d_{max}$.

$$||g_{1st} - g_{2nd}||_2 > d_{min} \quad (1)$$

$$||f_i - g_{1st}||_2 \leq d_{max} \quad (2)$$

Both of these constraints imposed by equations (1) and (2) have been discussed by Lowe [2]. First constraint discards ambiguous matches and works very well in practice. The constraint specified in equation (2) ensures that only accurate matches are taken into account. Although Lowe points out that this is not helping with SIFT descriptor matching we found out empirically that it improves results in our case significantly.

We can use the text-line information obtained during the preprocessing step to put further constraints on the matching process. Some samples of extracted text-lines and matched correspondences are shown in Figure 2. The features of a document image containing $n$ text-lines can be divided into $n$ bins such that each feature of a bin is closest to the bin's corresponding text-line. This allows us to divide the original matching problem into $n$ matching sub-problems of reduced size, which both improves efficiency and reliability of the procedure.

The matches are cleaned by imposing the epipolar constraint [18] which is depicted as follows:

$$p_r^T F p_l = 0 \quad (3)$$

where $p_l$ is a point from the left image and $p_r$ its corresponding match in the right image and $F$ id the fundamental matrix. $F$ is estimated from our set of matches by applying RANSAC [19].

By splitting up the original problem its computational complexity is reduced significantly. An average document image for example, as captured for our experiments, consists of about forty text-lines and performing feature detection on it produces about forty thousand features. Assuming linear search, which should roughly approximate exact NN methods in our case, this means the algorithm has to perform $n_{naive} = 40000^2 = 1.6 \times 10^9$ distance measurements. When restricting matching by text-lines the same process can on average be performed with $n_{restricted} = 40 \times 1000^2 = 4 \times 10^7$ comparisons. We can see that the number of required operations has been reduced to $n_{restricted}/n_{naive} = 1/40$, i.e. linearly in the number of text-lines, making practical application actually feasible as shown in our experimental results.

## V. Experimental Results

Feature extraction was carried out using the OpenCV implementation of SURF with 128-element descriptors, a Hessian threshold of 200, five octaves and eight layers per octave. The constraining parameters for matching regarding Equations (1) and (2) were set to $d_{min} = 0.05$ and $d_{max} = 0.3$. For comparison purposes the experiments are performed with and without using text-lines and are evaluated for both exact nearest neighbor search using kd-tree and FLANN[2].

The first column in Table I depicts the method used for matching along with the text-line constraint used. Second , third and fourth column show the number of matches found, number of correct matches and the percentage of correct matches respectively averaged over the whole dataset. The last column in Table I depicts the required average run-time for feature matching. The average is computed as follows:

$$p_{avg} = \frac{1}{n} \sum_{j=1}^{n} p_{I_j}, \quad p \in \{m, c, t\} \quad , I_j \in I = \{I_1, I_2, ..., I_n\}$$

where $p_{I_j}$ depicts the attribute taken from the j-th image $I_j$ from the dataset $I$. The symbols $m$, $c$, $t$ in attribute set represent the total number of matches, correct matches and time respectively. The last column in Table 1 shows that matching time has been significantly improved using text-line information. Hence providing evidence for the first claim that our method is efficient.

To measure the improved reliability of feature matching, we consider the percentage of correct matches remaining after RANSAC. A fundamental matrix will be computed using available matches and outliers are removed which does not satisfy the epipolar constraint as given in Equation 3. In our experiments, the probability that the fundamental matrix is correct given the matches is set to 0.99 and the maximum distance of a point from its corresponding epipolar line is set to one pixel.

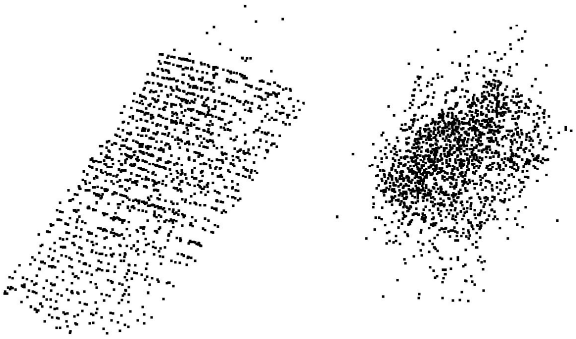


Figure 3. 3D reconstruction of book page without RANSAC. Left side shows the proposed method. Right side shows default matching.

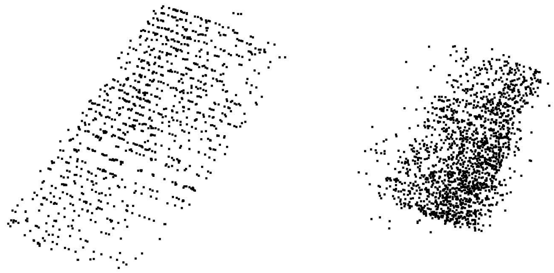


Figure 4. 3D reconstruction of book page with RANSAC. Left side shows the proposed method. Right side shows default matching.

The fourth column of Table 1 shows that using text-line information decreases the number of outliers. This will result in a better approximation of the fundamental matrix. This effect can be observed in the 3D reconstruction of a sample document image with and without use of text-line information which has been shown in Figure 3. No cleanup has been performed. It can be seen clearly that using text line information produces an almost accurate 3D model with good coverage of book surface. Figure 4 shows the same models from Figure 3 but cleaned up with RANSAC. There is almost no change in the 3D model reconstructed using text-line correspondences, but the unrestricted model still requires filtering before it could be used for any practical purpose.

## VI. Conclusion and Future Work

An approach for improving efficiency and finding reliable matches for 3D reconstruction for document images has been presented, which is based on matching constraints using document layout information, specifically text-lines. The process consists of preprocessing, including binarization, text-line segmentation, application-specific labeling and establishment of text-line correspondences. Robust features are detected, grouped by text-line and matched later on.

[2] M. Muja, Flann, fast library for approximate nearest neighbors, 2011, http://mloss.org/software/view/143/

Table I
ROBUSTNESS AND EFFICIENCY MEASURED BY PERCENTAGE OF CORRECT MATCHES AND AVERAGE RUN-TIME EVALUATED OVER WHOLE DATASET

| Method | Total Matches | Correct Matches | % | Time(sec) |
|---|---|---|---|---|
| KD without text-lines | 3281.62 | 1001.25 | 29.98 | 16803.17 |
| FLANN without text-lines | 2046.12 | 514.40 | 24.64 | 80.28 |
| KD with text-lines | 1946.23 | 1327.78 | 67.75 | 101.36 |
| FLANN with text-lines | 5083.34 | 2460.25 | 44.61 | 20.81 |

Experimental results are reported, which show the effectiveness of the proposed method. The method can be improved in various way: a robust strategy for finding better text-line correspondences is a possible enhancement to the naive approach used in this paper. Also the usage of specialized feature descriptors for text could be used for improving performance. Including extra constraints e.g limiting the matches by the maximum disparity value could also increase the performance of feature mapping. A GPU implementation of the proposed approach could result in real time matching performance due to the fact that features belonging to each grouping entity can be matched independently.

## ACKNOWLEDGMENT

The authors would like to thank Doğan Can Şen for capturing the document camera images that were used in this paper's experimental evaluations.

## REFERENCES

[1] 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, 2004. ICPR 2004.*, 2004, pp. 482–485 Vol.1.

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[4] S. S. Bukhari, F. Shafait, T. M. Breuel, "Coupled snakelets for curled text-line segmentation from warped document images (accepted for publication)," *International Journal on Document Analysis and Recognition*, 2011.

[5] S. S. Bukhari, F. Shafait, and T. M. 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. ICDAR*, 2011.

[6] V. Venkateswar and R. Chellappa, "Hierarchical stereo and motion correspondence using feature groupings," *International Journal of Computer Vision*, vol. 15, no. 3, pp. 245–269, Jul. 1995.

[7] R. Horaud and T. Skordas, "Stereo correspondence through feature grouping and maximal cliques," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 11, pp. 1168–1180, 1989.

[8] P. Moallem and K. Faez, "Fast edge-based stereo matching algorithm based on search space reduction," in *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*. IEEE, 2002, pp. 587–596.

[9] J. Kim, H. I. Koo, and N. I. Cho, "Camera-based document digitization using multiple images," in *2008 15th IEEE International Conference on Image Processing ICPR 2008*. IEEE, 2008, pp. 1025–1028.

[10] M. Pilu, "Undoing paper curl distortion using applicable surfaces," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE Comput. Soc, 2001, pp. I–67–I–72.

[11] A. Ulges, C. H. Lampert, and T. Breuel, "Document capture using stereo vision," in *Proceedings of the 2004 ACM symposium on Document engineering - DocEng '04*. New York, USA: ACM Press, 2004, p. 198.

[12] F. Shafait, J. van Beusekom, D. Keysers, and T. M. Breuel, "Document cleanup using page frame detection," *International Journal on Document Analysis and Recognition*, vol. 11, pp. 81–96, October 2008.

[13] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[14] F. Shafait, D. Keysers, and T. M. Breuel, "Efficient implementation of local adaptive thresholding techniques using integral images." in *15th Document Recognition and Retrieval Conference (DRR-2008)*, 2008.

[15] F. Shafait, D. Keysers, and T. Breuel, "Performance evaluation and benchmarking of six-page segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 941–954, June 2008.

[16] R. Weber, H.-J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *Proceedings of the 24rd International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., 1998, pp. 194–205.

[17] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009, pp. 331–340.

[18] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.