Perspective Correction of Panoramic Images created by Parallel Motion Stitching

João Gonçalves
Fraunhofer AICOS
Rua Alfredo Allen 455
4200-135 Porto,
PORTUGAL
joao.goncalves@fraunhofer.pt

David Ribeiro
Fraunhofer AICOS
Rua Alfredo Allen 455
4200-135 Porto,
PORTUGAL
david.ribeiro@fraunhofer.pt

Filipe Soares
Fraunhofer AICOS
Rua Alfredo Allen 455
4200-135 Porto,
PORTUGAL
filipe.soares@fraunhofer.pt

ABSTRACT

This paper deals with the problem of correcting the perspective distortion in panoramic images created by parallel motion stitching. The distortion is revealed by lines that appear to converge at the infinity, but are actually parallel. A camera cart shoots from multi-viewpoints aiming a parallel motion to the scene that is photographed. The perspective effect arises on panoramas while stitching several images taken from the camera, slightly panning in both directions between shots along the motion path. In this paper, we propose a solution to handle different camera translation motions and be able to stitch together images with a high-level of similarity, also having repetition patterns along a vast continuity of elements belonging to the scene. The experimental tests were performed with real data obtained from supermarket shelves, with the goal of maintaining the correct amount of product items on the resulting panorama. After applying the perspective correction in the input images, to reduce cumulative registration errors during stitching, it is possible to extract more information about the similarity between consecutive images so that matching mistakes are minimized.

Keywords

Affine transformation, panorama, stitching, parallel motion, multi-viewpoint, similarity, retail.

1 INTRODUCTION

Image stitching can be a complex sequence of image processing steps, especially when considering stitching several high resolution images photographed with wide-angle or fisheye lenses, at close range from the scene. Every time a high-level of similarity occurs between a pair of images, and a repetition pattern exists with a vast continuity of elements belonging to the scene, the errors in the final panorama rapidly rise with the number of pictures to blend. This is particularly important in scenarios where the exact number of elements of the reality should remain in the captured panorama.

For retailers, keeping supermarket shelves stocked is a vital part of running a successful operation. Monitoring shelves have always been an expensive and inefficient manual process, requiring stock clerks to do it throughout the day. The aforementioned stitching problem gets worse if one needs to capture the shelves of a long aisle of a supermarket. There is a limited distance between two opposite shelves in an aisle. The approach for a single photograph from the opposite side would either, capture a short portion of the shelf or have distortion towards the edges for wider field-of-views (using a fisheye lens for instance). Stitching errors must be avoided because one shelf has a planned number of items of a given product, and the final panorama must have the same number of items for control.

The problems described are not exclusive to supermarkets. In general, single-perspective photographs are not very effective at conveying long and roughly planar scenes, such as a river bank or the facades of buildings along a street. In those cases taking a photograph from a faraway viewpoint could be an alternative, but usually it is not possible to get far enough in a dense city to capture such a photograph. Even if it was possible the result would lose the track of perspective depth.

The stitching technique proposed in this work should not be confused with classic stitching around a view point. Rather than relying on a single projection, Kopf et al. [Kop09] introduce locally-adapted projections, where the projection changes continuously across the field-of-view. They also employ a perspective correction while ensuring continuous transitions between captured regions, thus preserving the overall panoramic context similar to how humans perceive the scene. However, the user must specify where the lines are straight. In addition, Anguelov et al. [Ang10] use a more complex image acquisition with multiple cameras calibrated among each other, on a street view scenario which has distinct constrains from a feature-based homography. The resulting 360° panoramas do not provide a good visual perception of a larger aggregate such as a whole city block, as discussed in [Kop10]. Agarwala et al. [Aga06] introduced an approach of producing panoramas that displays these types of long scenes. However, the user must specify the dominant plan of the scene and do not take into account the similarity and repetition of elements in the photos, as the case of products in supermarket shelves that should not appear merged due to stitching errors. In addition, the camera is manually moved and positioned along the scene.

The present paper departs from [Aga06] by adding motion to the multi-viewpoint stitching problem, to build a parallel motion stitching of photographed shelves, fully automatically. A camera cart shoots from multi-viewpoints and a slight motion of the camera as a very significant impact on the final panorama. When the camera moves in parallel to the scene that is being photographed, the parallel field-of-view is correct from the center of one image to the center of the next image, but it is not correct to the final image plane that appears in perspective projection, i.e. the last image is in perspective projection in relation to the first image. In particular, we focus on auto-calibration on a single photo. It is known that the final convergence of multi-image stitching depends (like any problem of optimization) how close the initial parameters are to the optimal solution. We aim to build a high resolution and long panorama and, therefore, we need a method that makes all images globally consistent without major artifacts.

2 METHODOLOGIES

In this paper, a stitching pipeline is proposed to essentially combine the orientation information extracted from parallel lines in the images with perspective effect, with the epipolar geometry to remove artifacts on the panoramic images caused by the semi-translational motion (translation plus rotation) of the camera that is panned relative to the scene. Any orientation changes are even more prominent with fisheye lenses.

The novelty of the paper lies in: preprocessing the images individually to remove the affine transformation (perspective) using position and angle from parallel lines present on image; roughly compute the region with high similarity between consecutive images; and as a last step, iterative matching process optimized by the homography angle along a specific plan.

The present methodology aims to reduce the accumulation of errors in registration process (homography-based stitching) from stitching pipeline presented by [Bro07] which is implemented on OpenCV [OpenCV]. Brown and Lowe 2007 extract features from all images. Then using those pairs of features, a homography matrix is calculated. This homography matrix is then used to warp one of the images onto other one, correcting local matching mistakes by a global regularization of images without a defined order. In this work in which the images follow an order, these kind of mistakes are solved even before the global regularization step, making the initial guess to be closer to the final global homography estimation.

We aim to overcome some problems of feature-based approaches, as images with lack of texture or similarity of features, which in our case is very common. Since all their camera spaces lie in the same horizontal plane in world space, only the translation plane from the camera space are collinear between all shots and coplanar with the translation plane from world space. Moreover, the different camera spaces are roughly aligned with world space which is not enough to turn the final panorama flat and rectilinear.

Figure 1 presents the proposed stitching pipeline. First the photos with a fisheye lens are acquired in motion (see Figure 2). Then, compute homography based on information from the parallel lines and apply affine transformation to the input images. After this it is possible to find the most similar region between consecutive images, so that matching mistakes are avoided. Moreover we improve the RANSAC (Random Sample Consensus) [Fis81] convergence (less iterations) by using this information. After this preprocessing stages we are ready to finish the registration and composing the images that are taken from parallel motion. Further details on the stitching pipelines are presented in the next sections.

2.1 Defisheye

Fisheye lenses create hemispherical images that must be undistorted to create a linear panorama. To accomplish this we follow Devernay and Faugeras [Dev95a] approach. The process from distorted to undistorted the fisheye effect is reversed so that the correction of the whole images is a lot faster. From the radius of the undistorted image $\mathbf{r_u}$ (the distance in pixels from the center of the image) (1), the correspondent radius of the distorted image $\mathbf{r_d}$ (2) is calculated where $\mathbf{f}$ is the apparent focal length of the fisheye lens, which is not necessarily the actual focal length of the fisheye. Basically, for each pixel in the processed image the corresponding pixel is determined in the distorted image.

$$r_u = \sqrt{x^2 + y^2} \tag{1}$$

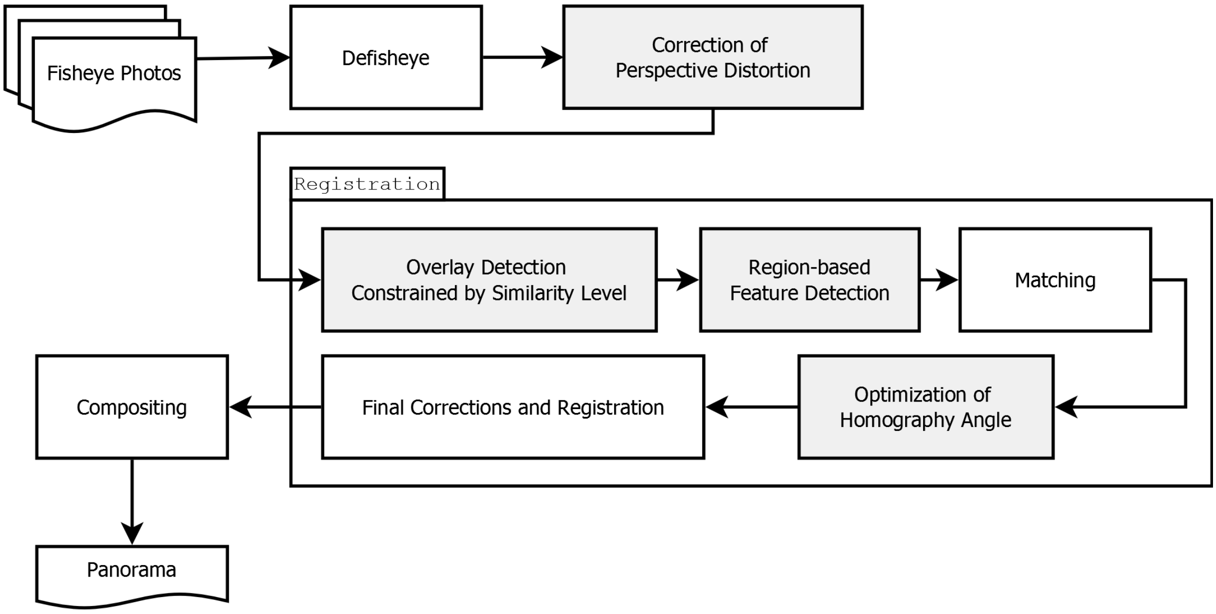

Figure 1: Stitching Pipeline. The main contributions are highlighted in gray.

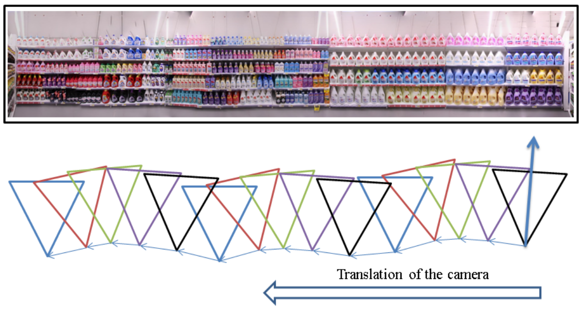

Figure 2: Example of scene to photograph.

$$r_d = f \arctan \frac{r_u}{f} \tag{2}$$

2.2 Correction of Perspective Distortion

Straight lines are common in man made environments, especially in supermarket aisles. Devernay and Faugeras [Dev01b] introduced the idea of "Straight lines have to be straight" to calibrate intrinsic camera parameters for a set of cameras. A similar approach is employed here, to compute the affine transformation of the image and posterior warp, but making the orthogonal plane from camera space roughly aligned between shots along the motion path of a camera cart. From images with parallel lines we can compute the orientation of the camera with respect to the scene. Considering a set of coplanar parallel lines in 3D world space like the ones in Figure 3. The lines in the image appear to converge when the camera is panned. Due to that, we search lines where the convergence is higher in the top and bottom of the images (lines closer to the center of the camera remain more straight) where it is more probable to find a line with the biggest slope. Using a line detector algorithm, we search for these lines (shelves) and calculate the slope that makes them converging, and then warp the image so they do not appear to be converging.

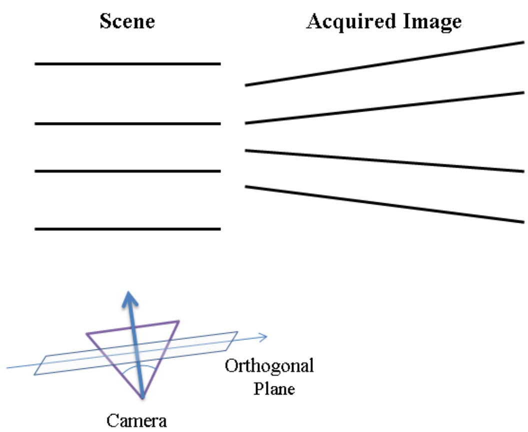

Figure 3: Example of panning camera. Left: the scene to be captured. Right: the image obtained due to the orthogonal plane not being parallel to the scene.

The perspective distortion is revealed by lines that in camera appear to converge at the infinity, but are actually parallel. This effect is based primarily on the panning and distance from the camera to the scene and, to some extent, on the focal length of the lens. It appears exaggerated with a fisheye lens because a bigger field-of-view is captured. In case the camera is panned, one half of the image is covering more area in 3D space than the other half, resulting in the convergence of lines in the photo (see Figure 3).

 ISBN 978-80-86943-66-4

The classical Hough [Dud72] transform provides a powerful and robust technique for detecting these lines, other curves or predefined shapes in a binary image. Our objective is to locate nearly linear arrangements of disconnected white spots (most probably "broken" lines). Considering that a straight line in the input image is defined by the equation (3) (polar coordinates), where

$$h : (x,y) \rightarrow \rho = x * cos(\theta) + y * sin(\theta) \qquad (3)$$

ρ and θ are two unknown parameters whose values have to be found. Every point in the binary image has a corresponding ρ and θ. Actually, these points form a sinusoidal curve in (ρ,θ) space called Hough Transform (HT) image. The whole HT image form a multitude of overlapping sinusoids that may converge on the same spot. The (ρ,θ) address of such point indicates the slope θ and position ρ of a straight line. This information is used to estimate the homography on a single image.

The scene presented in Figure 3 contains more than one possible line. Even if we divide the image on two halfs we can estimate the address of two lines (four points: two on the top half and two on the bottom half). To compute the homography for correcting the affine transformation on the image, eight points are needed: four points that define the correct image and four points to define the warp image.

2.3 Registration

Image registration is the process to assign the different coordinate system of every image that compose the final panorama, to a single and unique coordinate system. Stitching by feature-based homography requires registration to determine the overlapping region, i.e. to compute the homographies from one image to the subsequent. The key problem in image registration is to find a conversely relation from one image to another especially on perspective image. So first we need to know the coordinate system for every consecutive pair of images. For that task we use the Oriented FAST and Rotated BRIEF (ORB) [Rub11] to find features in every image that compose the panorama. The ORB descriptors are computed from those features, aiming a description of a point that is unique as possible, and matching the most similar descriptions according to Lowe et al. [Low04] to form a panorama, from which the information is extracted from the set of overlapping regions. That information is necessary to estimate the final corrections and obtain the resulting estimated homography (global coordinate system).

2.3.1 Overlay Detection

After the previous step of correcting perspective projection, the consecutive images should be aligned. Using the L2-norm, we compute the regions that are more similar inside the overlap formed between each pair of consecutive images. This step is taken in conjunction with a vertical region-based feature detection (detailed in the next section), as we pretend to avoid matching mistakes influenced by a high repetition of products on supermarket shelves. A mask of descriptors can be created knowing the position of this region on the image, which means that we discard the matches that are outside of this region. In case the estimated homography from this region are very far apart from the basic understanding of the scene, this region can slide.

2.3.2 Region-based feature detection

In stitching by feature-based homography on which image matching is to be performed, there are basic requirements that interesting points should have:

- i clear and mathematically well-founded definition (estimation of descriptors)
- ii well defined position on image space (detection of features)
- iii local image structures around the interest point, that are mostly unique if it were possible
- iv stability for a reliable computation of a interest point under image deformations and illumination variations (invariance across view points).

To estimate reliably the initial guess homography, the intra-image features must be distinctive from anything else in the same image (i,ii and iii). Since objects are likely to repeat, these invariant features can also appear repeated in the scene, and finding corresponding image points can be a huge challenge.

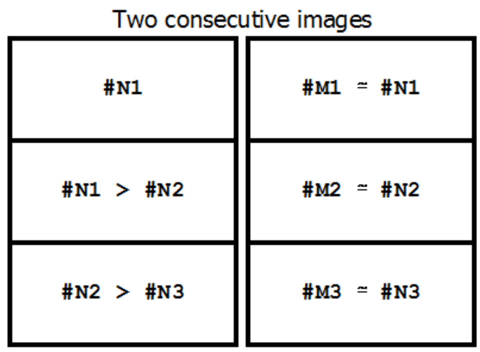

Figure 4: Diagram to extract different cardinality of features in the vertical direction of each image. N and M are detected features.

Knowing the most similar region already detected between consecutive images (horizontal), and because we are photographing shelves with a fixed scheme, it is possible to avoid matching mistakes by extracting different cardinality of features in the vertical direction.

 ISBN 978-80-86943-66-4

From top to bottom, the image is split into sections (see Figure 4). Similar features in the same image have different distance between neighbor features (since the cardinality is different), to improve the posterior matching process in the stitching pipeline.

2.3.3 Matching

The matching process consists in finding the image features that appear in other image with the same descriptors. In this work, the ORB descriptor is used to locate points of interest and the description of those points. Without the previous steps the number of ambiguity between matches is huge, partly due to the characteristics of the supermarket shelves. On the other hand, the stitching pipeline used in [Bro07] find this ambiguity after processing the initial guess resolving the auto-calibration problem (homography between consecutive images). Brown and Lowe, first do the features matching based on local description (nearest neighbor strategy) and label them with multiple matches as ambiguous. Then, in a second stage the ambiguous matches are removed imposing globally consistence constraint which is solved using the Levenberg-Marquardt algorithm.

2.3.4 Optimization of homography angle

To compute a reliable global homography, it is necessary that the initial guess is roughly estimated and close to the optimal solution which implies that the matching process do not have ambiguities, i.e., the images would not have similar structures or repeated patterns. Since this is not possible in supermarket shelves, by using the proposed methodology the first guesses are well estimated. Homographies that describe the affine transformation between pairs of images are estimated using RANSAC estimation method (the outliers of corresponding points can be removed more easily by the RANSAC), based on Singular Value Decomposition (SVD) to extract the rotations and translations matrix out of the estimated homography as explained in [Zis04].

The angle in the orthogonal plane from the rotation matrix is used as a measure for the amount of misregistration (as we move in parallel to the scene, the final stitching should be flat and rectilinear). If the resulting angle is above a threshold the descriptor masks are slided and the matching process is run again with different features. More detail on homography decomposition for extraction of Euler angles can be found in [Mal07].

2.4 Compositing

This stage follows the steps of the stitching pipeline in [Bro07]. A portion of different images is cut and pasted into the final panorama. Naturally, the simple cut and past step leaves artificial edges in the overlapping regions due the difference in camera gain and scene illumination.

Once we have registered all of the input images with respect to each other, we need to decide how to produce the final panorama image. This involves selecting a final compositing surface (flat, cylindrical, spherical, etc.) and view (reference image). It also involves selecting which pixels contribute to the final composition and how to optimally blend these pixels to minimize visible seams, blur, and ghosting [Sze11].

A common approach in warp several images to compose the final panorama is to use one image as reference and warp the others to the reference. With this approach the warp to the last image is the accumulation of all warps, which means that if the referenced image are in perspective projection the second image to be warped must continue the perspective projection in order to respect the global homography. The sequence of successive warps makes the final panorama appear shrinking or with expansion. This is the case when one tries to stitch images taken from parallel motion. For images that have large rotations it can not be handled. Our approach tries to reduce this effects on final panorama.

3 RESULTS AND DISCUSSION

The collection of images for the experimental test was obtained at a large supermarket of Sonae MCH, Portugal. The dataset consists of ten aisles with length varying between 6m and 13m, captured one meter from the supermarket shelves. The setup is focused on creating a high-quality panorama of complete grocery store aisle (see Figure 2). Due to the constructions of grocery stores the distance between aisles is too short to take a photo of the full aisle in length. The solution is to move the camera along the aisle and take several photos, stitching them at the end. Problems do arise from this approach: short distance implies using wide-angle or fisheye lens to capture the full height of aisle. The motion of the camera is not linear and it may appear panned relatively to the previous capture.

Spizhevoy and Eruhimov 2012 [Spi12] do not use pattern-based calibration. The auto-calibration of the camera is done using the epipolar geometry only, assuming that the camera mainly has rotation between viewpoints. Moreover, the work in [Spi12] has stronger constraints with respect to the minimal distance to scene, that has to be twice the value employed here. Contrarily, our method does not require knowledge of the intrinsic camera parameters.

Perfect final panorama implies that all the photos are taken exactly parallel to the aisle, but since the camera is in motion it is more difficult to maintain it perfectly aligned. This means that the photos need to have a large overlapping region (typically at least 50%). Due to the

 ISBN 978-80-86943-66-4

Algorithm 1: STITCHING IMAGES FROM PARALLEL MOTION

Input: A sequence of n ordered images

for $i = 1 \to n$ **do**

1. Undistort fisheye effect.
2. Compute hough space for top and bottom of image and extract the biggest line in these two regions.
3. Using the slope from the previous lines, compute homography and warp the image to undo the convergence of lines.

end

for $j = 2 \to n$ **do**

4. Compute mask for high similar region between consecutive images using norm L2.
5. Extract ORB features for image j respecting cardinality constraint.

while *angle* $>$ *threshold* **do**

6. Find k nearest-neighbours for consecutive features.
7. Find geometrically consistent feature matches using RANSAC, in combination with masks calculated before, to solve the homography between images pairs.
8. Verify image matches using the angle in the orthogonal plane, if false slid masks.

end

end

9. Perform bundle adjustment to solve the rotation θ_1 θ_2 θ_3[Mal07] and focal length for all cameras.
10. Render panorama using multi-band blending.

Output: Panorama image

nature of supermarket aisles, many times the products appear repeated continuously, so the probability of having similarity between two consecutive images is high.

To validate the quality of the correction method proposed herein, we compare two measures that reflect how far from linear a given panorama has become: the percentage difference between the height in the beginning and the end of the resulting panorama, as % Difference Height (DH); the global slope of the aisle perspective distortion, as % Global Slope (GS). The results are summarized in Table 1 that reflects the improvements on different aisle panoramas. The best results obtained with [Bro07] (Before) are compared with our approach to the problem (After). The improvements noted by lower values of DH or GS occur mainly because of the perspective correction, overlay detection method and homography angle refinement. Also, a relevant step is the extraction of different cardinality of features in the vertical direction in the individual images.

Figure 5 and 6 show the gain in visual quality, not only in % GS that represents the degree of perspective effect, but also the DH that shows how misestimated the panorama is compared to the real scene. From Figure 7 to 8, we can visualize how the missing of product items was recovered. As observed in Figure 9, the most perceptible artifact in the final image is the parallax effect in rear products. This parallax level is acceptable for our goal, since only the front facings of the products count. To improve this kind of artifact an alternative would be to follow Zheng et al. 2011 [Zhe11]. The authors construct panoramas from multi-view points using epipolar geometry to remove the perspective effect on final panorama, making use of optical flow to avoid parallax effects. This method requires video capture which has a lot more information to compute a reliable epipolar geometry.

4 CONCLUSION

Parallel motion stitching is definitely an interesting technique for panorama production. This article proposes a stitching pipeline for motion photography, using perspective correction, similarity-constrained overlay detection and homography angle optimization. The results allow us to conclude that it is possible to correct major artifacts in stitched images of supermarket shelves, that usually require a high level of visual quality in the resulting panorama. It is important that the final image avoids a lack of elements in the scene compared to the reality, if possible resembling a unique capture of the whole scene. The panoramas exhibit detailed information even with a photo shooting at multi-view points, without worries on limited spaces, and also during motion. The goal proposed was accomplished without manual intervention on capture nor image processing.

Table 1: Correction quality in parallel motion stitching measured by % Difference Height (DH) and % Global Slope (GS). Before: The best results obtained with [Bro07]. After: Our approach. Lower is better.

	Before		After	
	%DH	%GS	%DH	%GS
Aisle 1	50.45	5.35	6.02	1.43
Aisle 2	38.70	5.49	5.18	1.80
Aisle 3	42.66	4.95	4.87	1.45
Aisle 4	38.37	6.84	2.50	1.82
Aisle 5	38.35	3.86	11.43	1.04
Aisle 6	41.66	4.87	10.45	1.11
Aisle 7	44.18	3.88	13.14	0.83
Aisle 8	41.17	4.24	6.67	1.05
Aisle 9	41.73	4.35	3.22	3.40
Aisle 10	44.54	3.51	12.55	0.95
Mean	41.70	4.47	5.60	1.08

5 ACKNOWLEDGMENTS

A special acknowledgment to Sonae MCH, Portugal for providing the environment for data collection. The authors acknowledge the financial support obtained from the European Regional Development Fund (ERDF) through COMPETE - Programa Operacional Factores de Competitividade (POFC), for the project ShopView (project number 30335).

6 REFERENCES

[Kop09] J. Kopf, D. Lischinski, O. Deussen, D. Cohen-Or, and M. Cohen, "Locally adapted projections to reduce panorama distortions," in Computer Graphics Forum, 2009, vol. 28, pp.1083-1089.

[Ang10] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google Street View: Capturing the World at Street Level," Computer, vol. 43, no. 6, pp. 32-38, Jun. 2010.

[Kop10] J. Kopf, B. Chen, R. Szeliski, and M. Cohen, "Street Slide: Browsing Street Level Imagery," in ACM SIGGRAPH 2010 Papers, New York, USA, 2010, pp. 96:1-96:8.

[Aga06] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski, "Photographing long scenes with multi-viewpoint panoramas," in ACM Transactions on Graphics (TOG), 2006, vol. 25, pp. 853-861.

[Bro07] M. Brown, e D. G. Lowe. "Automatic Panoramic Image Stitching Using Invariant Features," International Journal of Computer Vision, vol. 74, no. 1, pp. 59-73, Aug. 2007.

[OpenCV] OpenCV documentation site: `http://docs.opencv.org/index.html`

[Low04] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International journal of computer vision, vol. 60, no. 2, pp. 91-110, 2004.

[Wan08] Y. Wan and Z. Miao, "Automatic panorama image mosaic and ghost eliminating," in 2008 IEEE International Conference on Multimedia and Expo, 2008, pp. 945-948.

[Zis04] R. Hartley and O. Zisserman, "Multiple View Geometry in Computer Vision", second ed. Cambridge Univ. Press, 2004

[Dev95a] F. Devernay and O. D. Faugeras, "Automatic calibration and removal of distortion from scenes of structured environments," in SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation, 1995, pp. 62-72.

[Dev01b] F. Devernay and O. Faugeras, "Straight lines have to be straight,"Machine vision and applications, vol. 13, no. 1, pp. 14-24, 2001.

[Dud72] Duda, R. O. and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," Comm. ACM, Vol. 15,1972, pp. 11-15.

[Per10] A. Pernek and L. Hajder, "Perspective Reconstruction and Camera Auto-Calibration as Rectangular Polynomial Eigenvalue Problem," in 2010 20th International Conference on Pattern Recognition (ICPR), 2010, pp. 49-52.

[Mal07] E. Malis, M. Vargas, and others, Deeper understanding of the homography decomposition for vision-based control, 2007.

[Rub11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in Computer Vision (ICCV), 2011 IEEE International Conference on, 2011, pp. 2564-2571.

[Fis81] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Commun. ACM, vol. 24, no. 6, pp. 381-395, Jun. 1981.

[Sze11] R. Szeliski, Image stitching, in Computer Vision, Springer, 2011, pp. 375-408.

[Spi12] A. Spizhevoy and V. Eruhimov, "Problem of auto-calibration in image mosaicing," presented at the International Conference on Computer Graphics and Vision, GraphiCon, Conference Proceedings, 2012, pp. 27-32.

[Zhe11] E. Zheng, R. Raguram, P. Fite-Georgel, and J.-M. Frahm, "Efficient Generation of Multi-perspective Panoramas," in 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2011, pp. 86-92.

Figure 5: Panorama create with method from [Bro07] with uncorrected perspective effect.

Figure 6: Panorama create with proposed method of the same aisle in Figure 5, with correction.

Figure 7: Panorama create with method from [Bro07] with missing products.

Figure 8: Panorama create with proposed method of the same aisle in Figure 7, with correction and right number of product items.

Figure 9: Another example of panorama create with proposed method with correction.

 ISBN 978-80-86943-66-4