AI Guided Panoramic Image Reconstruction

Arnaud SCHENKEL, Laboratories of Image Synthesis and Analysis / PANORAMA, ULB, Belgium
Zheng ZHANG, Ecole polytechnique de Bruxelles, ULB, Belgium
Olivier DEBEIR, Laboratories of Image Synthesis and Analysis / PANORAMA, ULB, Belgium

Abstract: Due to the development of virtual reality, virtual tours have become commonplace. Panoramic image is the core to generate such immersive environments. However, the moving foreground will degrade the visual quality and cause ghosting artifacts on the panorama, making difficult the representations of most popular crowned city places as the Grand Place of Brussels where there are always people all the day. This paper addresses this problem by a novel panoramic image reconstruction pipeline, adding a moving foreground objects removal step, before a state-of-the-art stitching phase, to both guide the pictures acquisition and obtain a clean panorama. The proposed solution consists of analyzing the pictures in parallel with their acquisition to know whether they should be taken again, considering that it is difficult to obtain an image without any moving object and that each new contribution makes it possible to obtain new small background areas. Several model-based and artificial intelligence-based foreground removal approaches are proposed and evaluated. Best results are obtained by using an object detection convolutional neural network, Mask R-CNN (He et al., 2017), to detect foreground objects and using mixture of gaussians background subtractor, MOG2 (Zivkovic, 2004), to detect shadow. The pure background is extracted from a series of images according to the foreground detection masks. Experimental results show that our panorama generation pipeline effectively removes the moving objects.

Keywords: *Moving Object Detection—Moving Object Removal—Panoramic Imaging—Image Stitching*

CHNT Reference: Schenkel, A., Zhang, Z., and Debeir, O. (2022). ‘AI Guided Panoramic Image Reconstruction’, in Börner, W., Rohland, H., Kral-Börner, C. and Karner, L. (eds.) *Proceedings of the 25th International Conference on Cultural Heritage and New Technologies, held online, November 2020*. Heidelberg: Propylaeum.
doi:10.11588/propylaeum.1045.c14500

Introduction

The panoramic image gives an extensive angle of view, which is widely used in immersive environment generation and virtual reality. Used to create virtual tours of museums, art events, or world heritage sites, 360 panoramic view must take care of human presence during the acquisition to avoid any disturbance in the final product.

The construction of the panoramic image is accomplished by stitching multiple photos. However, when the photos are captured in a high-traffic place, the moving objects will degrade the visual quality and cause ghosting artifacts on the panorama because of the moving foreground objects in the seam of two images. Furthermore, the pure-background panorama without the interference of foreground objects is preferred when the scene is displayed. This project aims to solve the acquisition and the moving foreground interference problem in crowded places where it is never possible to have such a museum of world heritage without visitor, like popular city places such as the Grand

Place of Brussels where there are always people, even late at night, and generate a high-quality panorama of pure background visitors (Fig. 1).

Our solution consists of adding a moving foreground objects removal step, before a state-of-the-art stitching phase, to both guide the pictures acquisition and obtain a clean panorama, preferably without moving objects, but with coherent objects if is not possible due to some acquisition limitations. Several model-based and artificial intelligence-based approaches are proposed and evaluated. Best results are obtained combining both approaches: Mask R-CNN for object detection and MOG2 for shadow detection. Result therefore currently depends on the acquisition time and shots to remove visitors.

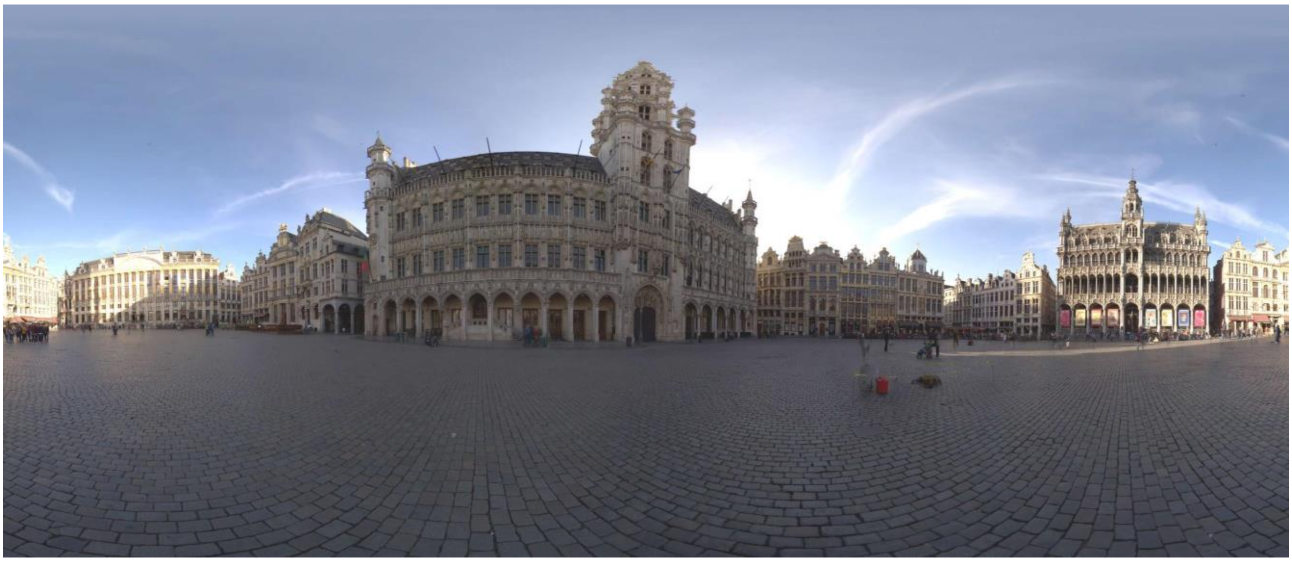

Fig. 1. Grand Place in Brussels, based on limited time acquisitions. Objects and visitors, stationary during all the acquisitions, could not be correctly removed (© Arnaud Schenkel).

Panoramic Photography

Since always, crowned and fortunate people liked paintings in general or panoramic view, carried out on walls, tapestries or panels, showing wars or hunting scenes. Its emergence among a wider audience dates more from the 18–19th century. The panoramic views then allow the public to be immersed in new surroundings, such as simple landscapes, but also topographical views and historical events, with paintings such as the Panorama of London from Albion Mills (1792), the Cyclorama of Jerusalem (1895), the Panorama of the Battle of Waterloo (1912), and the Pantheon of the World War (1914), or with pieces of art, like the Eidophusikon (1781, Philip de Loutherbourg), and the Diorama Theatre (1821, Louis-Jacques Daguerre and Charles Marie Bouton). Table 1 gives a summary of the characteristics of four well-known panoramic paintings.

Table 1. Characteristics of four well-known panoramic paintings.

Date	Name	Artist(s)	Dimensions
1792	London from the Roof of Albion Mills	Robert Barker	43 cm by 330 cm
1895	Cyclorama of Jerusalem	Paul Philippoteaux (based on Bruno Piglhein’s work	14 m high / 110 m in circumference
1912	Panorama of the Battle of Waterloo	Louis Dumoulin	12 m high / 110 m in circumference
1914	Pantheon of the World War	Pierre Carrier-Belleuse and Auguste Gorguet	14 m high / 123 m in circumference

The birth of photography did not profoundly change the style of panoramas, as they often remained a succession of images rather than a continuous image. From the beginning of the 19th century, specialized panoramic camera designs were being patented and manufactured for making panoramas. Several approaches are thus developed:

- using a specialized rotating lens camera and a curved filming plate: Joseph Puchberger invented the first-hand crank driven swing lens panoramic camera in 1843 but limited to record a 150-degree field of view instead of a full 360-degree view. In 1845, Frédéric Martens take panoramic shots of Paris using a chamber with a hundred fifty degrees opening;
- using a specialized camera with a wide field of view, up to 180°: Kodak proposed his first panoramic camera, the No. 4 Panoram, in 1899, to obtain quite similar result but with an easier solution;
- taking a series of images which were then shown placed next to each other to create one image.

Among the oldest and most famous panoramic pictures, Martin Behrman's eleven-panel panorama shows the state of San Francisco from Rincon Hill in 1851. Bernard Otto Holtermann and Charles Bayliss capture Sydney Harbor from Lavender Bay to produce one of the most impressive photographic achievements in 1875, composed of twenty-three albumen silver photographs (178 centimeters). Another example is the panorama of Verdun, filmed from the Fort de la Chaume in 1917.

Since the invention of digital photography, it is easier and much less expensive. But modern solutions are based on the same ideas. Fig. 2 gives a summary of the acquisition solutions to produce panoramic images.

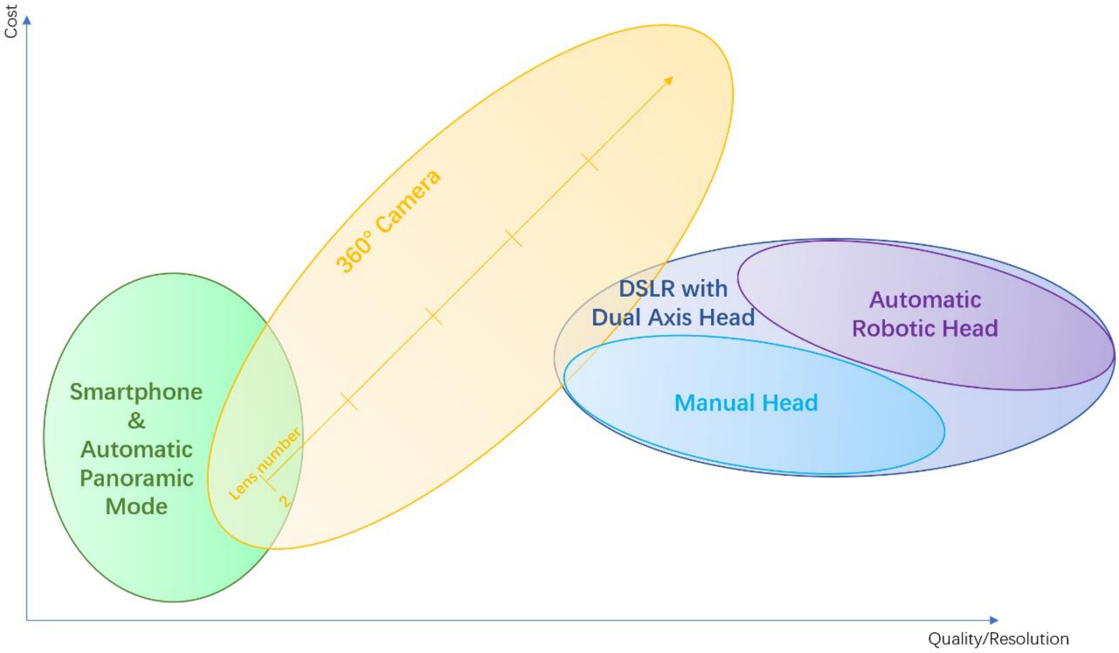

Fig. 2. Inventory of acquisition solutions for producing panoramic images (© Arnaud Schenkel).

Software solutions embedded in modern cameras or smartphones allow panoramic photographs to be taken simply by rotating the device. The result obtained is often limited in terms of resolutions and makes it difficult to deal with the presence of obstacles in the scene.

The 360° cameras have the main advantage of capturing the whole space at one time. However affordable 360° cameras are limited to two wide angle cameras, producing images of limited quality in terms of distortion and resolution; while devices composed of a larger number of cameras allow higher quality, also at a higher cost.

The current most common method for producing high details panoramas is to take a series of pictures by turning the camera between each shot, considering an overlap between two consecutive shots, covering the full spherical environment and stitch them together. Two solutions for that: manual acquisition or using automatic robotic head. The number and the angular positions of shots thus depend on the characteristics of the camera (sensor size, orientation), the lens type (rectilinear or fish-eye, and the focal length), the coverage and the overlap ratio.

Panoramic Image Stitching

The desired output is a 360-degree panorama, which is an immersive image containing information of the full enclosed-sphere scene from one viewpoint inside the sphere. The state-of-the-art stitching pipeline contains three phases: pre-processing, registration, fusion. Camera distortions and colors correction are corrected in the pre-processing phase. In the registration phase, the mapping between the pixels in two adjacent images is defined by a homography matrix, which is calculated by the estimated camera parameters. Homography matrix is generally computed based on pairs of matched points, extracted from the images using a feature detection algorithm and matched according to a matcher method. The estimated homography matrix is pairwise, which makes each homography is independent of others. If the images are stitched from the first to the last, the errors will accumulate, which causes lower visual quality. Therefore, global refinement of the camera parameters that minimize the misregistration is needed, which is named bundle adjustment (Szeliski, 2007; Chen et al., 2019). Knowing the camera parameters, the transformations between each pair of images are known. The images are then projected to a sphere according to the camera parameters, and then the equirectangular projection is used to map the sphere coordinates to plane coordinates for visualization, which is named image warping (Szeliski and Shum, 1997). The warped images are finally blended to a 360-degree panorama in the fusion phase.

A large amount of work has looked specifically at one or another step in the process to improve it, according to different criteria of quality, acquisition or computation time. Among these studies, El Abbadi et al. (2021) brings together a series of comparative analyses of various methods for feature extraction (e.g. SIFT, SURF, KAZE, ORB), point matching and homography estimation. Multiple solution are also proposed to smooth the transition between images by using seam finding or blending approach (Szeliski, 2007; Herrmann et al., 2020).

The presence of moving objects or people in the acquisitions still leads to defects in the result: the presence of artifacts, occlusions, ghosting effects, able to ruin the final composition or to hide some important details in an architecture or artifacts in a museum room... Simple stitching processing of these images does not allow dealing with the whole problem, allowing only to solve the issues in the overlapping parts partially. The state-of-the-art solutions only deal with the problem by having a sufficient number of acquisitions at the risk of presenting some aberrations in poorly covered area (like tearing, ghosting, duplication of recognizable objects or people).

Method

The proposed solution consists of analyzing the pictures in parallel with their acquisition to know whether they should be taken again. In animated site contexts, on the one hand, it is difficult to obtain an image without any moving object; on the other hand, each new contribution makes it possible to obtain new small background areas. The combination of all of these areas, therefore, makes it possible to obtain a completely cleaned panoramic image.

Figure 3 gives an overview of the complete pipeline of the proposed solution.

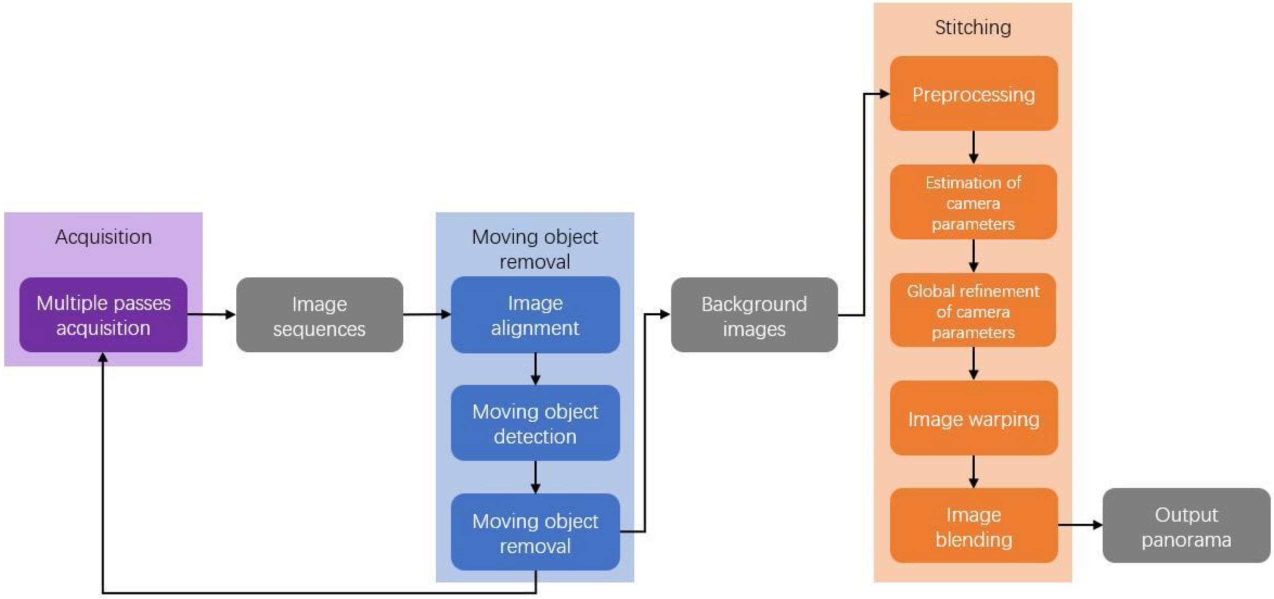

Fig. 3. The complete pipeline to generate panorama with foreground removal (© Zheng Zhang).

The three major steps are: acquisition, moving objects removal, and stitching.

1. Acquisition: datasets are captured pose by pose under control of the robotic head, using multiple passes when a picture contains moving objects, producing an image sequence for each camera pose.
2. Moving objects removal: Assuming that the pixels of the same coordinate are not covered by the foreground objects in all photos, the moving object removal approach can extract background information from the image sequences. The presence of a foreground object in the same picture's part in all the sequences implies the need for a new acquisition.
3. Stitching: After removing the moving foreground objects, there will be pure background images having overlapping parts, and then these images can be stitched to a panorama.

Moving Foreground Removal Approaches

Assuming that the pixels of the same coordinate are not covered by the foreground objects in all photos, the moving object removal approach can extract background information from the image sequences. The key idea is filtering out the foreground pixels and merging the remaining background pixels for each spatial point in the result. Therefore, the pixels at the coordinates (x, y) of each image captured with the same camera pose should correspond to the same spatial point, and the images in the sequence need to be aligned by homography transformation as preprocessing.

The classical approach is the median of images. However, this approach has ghosting artifacts in the area where the foreground objects are denser. A better approach is the foreground mask-based approach to remove the foreground. Some definitions of variables are given as follows for clarity: given an aligned image sequence $I_1, I_2, \dots I_i, \dots I_n$, $C_{c,i}(x,y)$ is the value of channel c of the pixel at coordinates (x,y) in image I_i. $C_c(x,y)$ is the value of channel c of the pixel at coordinates (x,y) in the result.

The workflow (Fig. 4) starts with foreground object detection, whose objective is obtaining a binary mask M_i that marks foreground objects for each image I_i in the sequence. $M_i(x,y)$ is the Boolean value at (x,y) in the mask M_i. The set $M_{x,y}$ contains the index of images in which the pixel at (x,y) is the background, i.e., $i \in M_{x.y}$ if $M_i(x,y) = False$. $N_{c,x,y} = \{C_{c,i}(x,y) \mid i \in M_{x,y}\}$, which is the set of values of channel c of unmarked (background) pixels at coordinates (x,y). The background is computed as:

$$C_c(x,y) = \begin{cases} median(\{C_{c,i}(x,y) \mid i = 1, \dots, n\}), & if N_{c,x,y} = \emptyset \\ median(N_{c,x,y}), & otherwise \end{cases}$$

Thus, the pixel values in the output are the median of unmarked pixels at the corresponding coordinates.

This approach can be summarized as extracting all background pixels from the sequence of images; then, the extracted pixels are merged to generate the result. If no background pixels are detected at a place, the value of that pixel is the median of all pixels at that position. Another solution to deal with such presence of a foreground object in the same picture's part in all the sequences is executing a new acquisition at that camera pose.

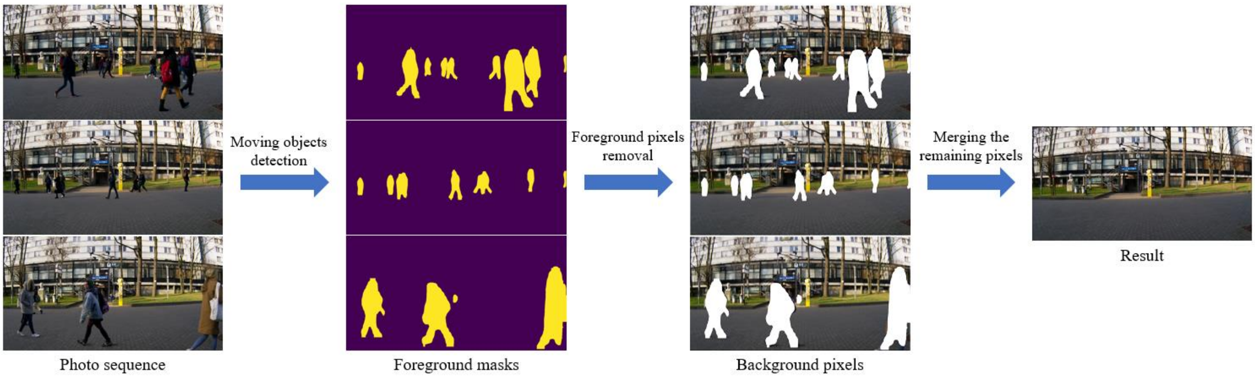

Fig. 4. Workflow of foreground-mask based approach: 1. Moving foreground detection; 2. Foreground masks generation; 3. Pixel merging (© Zheng Zhang).

Depending on the opportunities of acquisition; several scenarios are possible:

- the combination of the images allows to correctly and completely remove the foreground objects; the process then continues with the next step;
- the images acquired are insufficient to treat the problem: either it is possible to make new acquisitions, a new cycle of acquisition and foreground removal is carried out; either new acquisition is not possible (due to limitations of the acquisition time, for example, depends on functional or lighting conditions), an image is then selected to present an entire obstacle in order to give a tangible result.

Moving Foreground Objects Detection

Several foreground detection methods are implemented: Three-frame differencing, MOG2, KNN, YOLO, and Mask R-CNN. Three-frame differencing, proposed by Kameda and Minoh (1996) detects moving objects in an image sequence by calculating the difference between two images, considering a third allows to identify the persistent elements, and therefore differentiate the foreground and the background. Because the proposed approach is executed on discrete images, the three frames do not have to be in chronological order as the original version. The schematic diagram of the novel proposed three-channel version is shown in Figure 5. Assuming there are n photos in the sequence, I_i describes the $i-th$ image. The masks of the first image I_1 and the mask of the last image I_n are generated from I_n, I_1, I_2 and I_{n-1}, I_n, I_1, respectively. Instead of changing the image to grayscale as the original version, the proposed three-frame differencing implementation is applied separately on three channels; a binary closing operation follows the logical 'AND' to reduce the holes, and then the binary masks of the three channels are combined by the 'MAX' operation. If a pixel is marked as a pixel of moving foreground on one channel, the pixel will be marked in the final result.

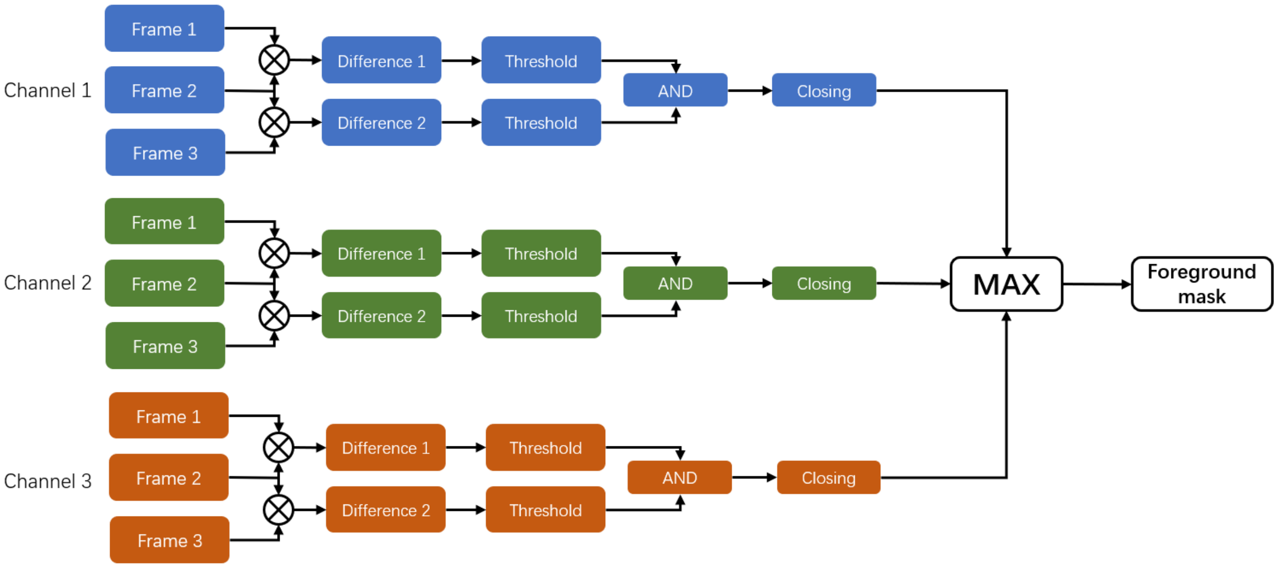

Fig. 5. Schematic diagram of multichannel three-frame differencing (© Zheng Zhang).

The MOG2 (Mixture of Gaussians v2, background subtractor with shadow detection described by Zivkovic (2004)) and KNN (K-nearest neighbors-based foreground segmentation described by Zivkovic and Van der Heijden (2006)) are background modeling methods usually applied to videos and need hundreds-frames initialization. MOG2 is a background subtractor with shadow detection. This approach is a Gaussian Mixture Model (GMM) based algorithm, which builds a statistical background model of the scene. The image to be detected is compared with the background model, and the pixels that do not fit the background model is marked as pixels of a foreground object. The KNN algorithm maintains a fixed-length memory to store the values of the pixels in a period T for each pixel. For a newly arrived frame, if the value of the pixel has more than k neighbors within its neighborhood of radius r in the memory, which stores the historical values, the pixel at the newly arrived frame will be assigned to the background.

Both MOG2 and KNN are designed for processing the consecutive frames in a video, and these model-based approaches usually need hundreds of frames to initialize the model. However, the used image sequence is not consecutive frames, so we need to initialize the background model artificially. Three fast initialization methods are proposed:

- using the median of the images in the sequence to initialize;
- initializing by a series of gamma adjustment of the median;
- pre-feeding the image sequence once, and then feed the image sequence again to obtain the masks.

The best initialization method of each of these three model-based approaches is different. The comparison of varying initializations will be given in the experiments section. The output foreground masks of MOG2 and KNN are followed by dilation operation and closing operation to obtain the optimal masks.

YOLO (You Only Look Once, real-time and end-to-end object detection convolutional neural network described by Redmon et al. [2016]) and Mask R-CNN (Region-based Convolutional Neural Network method with object mask prediction, described by He et al. [2017]) are deep neural networks-based approaches. These two object detection networks can detect foreground objects directly with knowing the class of the foreground objects, e.g., pedestrians and vehicles. YOLO is the most widely used real-time and end-to-end object detection convolutional neural network. It aims to detect multiple-categories objects in the input image. For each detected object, it returns the center, height, and width of the bounding box. The foreground mask is generated by setting the pixels inside the bounding boxes to be True. Object detection of the YOLO yields bounding boxes, so the foreground mask of YOLO is composed of rectangular. That will waste some background pixels around the foreground objects. Therefore, the Mask R-CNN, which can generate a segmentation mask, is brought to be compared with YOLO. Since YOLO and Mask R-CNN cannot detect the shadow of objects, a shadow mask is added to the object detection mask by identifying the shadow of moving objects using MOG2. The final result is obtained by binary 'OR' operation between the object detection mask and the additional shadow mask.

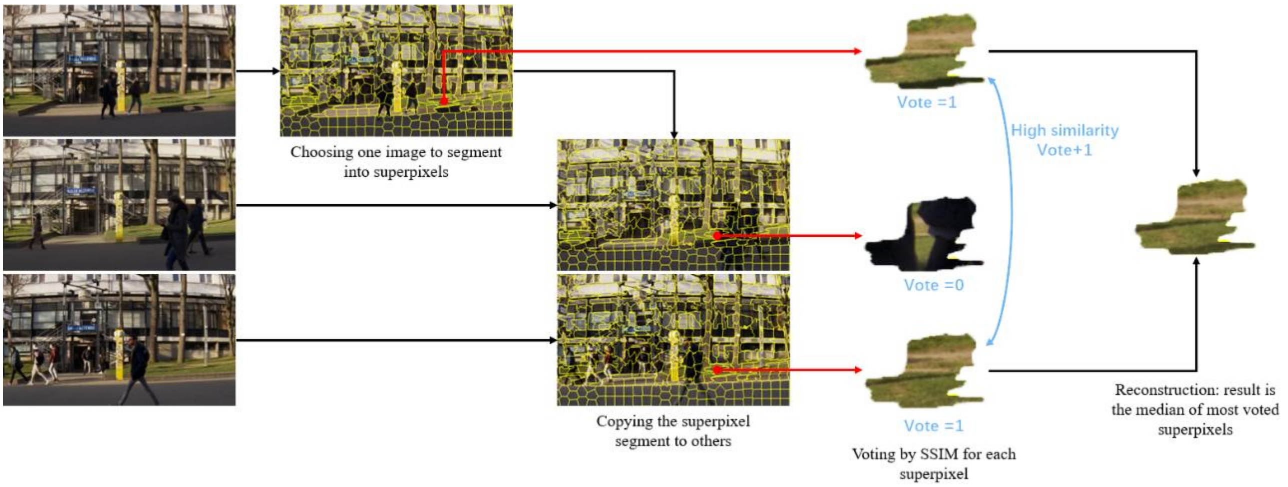
Choosing one image to segment into superpixels
Vote =1
High similarity Vote+1
Vote =0
Vote =1
Copying the superpixel segment to others
Voting by SSIM for each superpixel
Reconstruction: result is the median of most voted superpixels

Fig. 6. Superpixel voting approach. An image is segmented into superpixels and this segmentation is copied to others. Then resulting superpixel is generated from a cluster of similar superpixels (© Zheng Zhang).

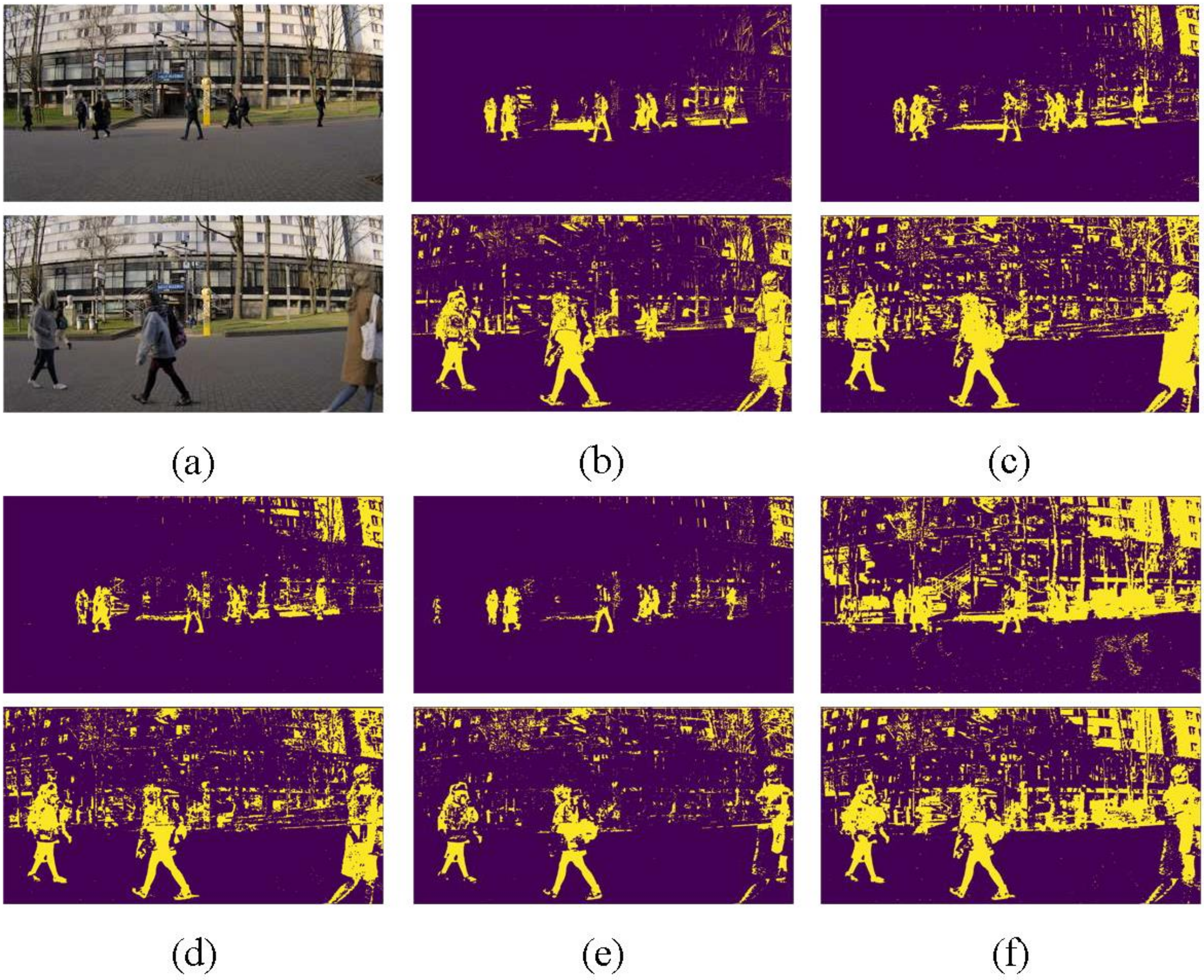

Fig. 7. Examples of obtained masks for two picture following described methods: (a) Training images, (b) Three-frame differencing, (c) MOG2 (median initialization), (d) MOG2 (training images initialization), (e) MOG2 (gamma initialization), (f) KNN (median initialization) (© Zheng Zhang).

Besides the foreground mask-based approaches, a superpixel-based method which is inspired by Su (2018) is created. Superpixel is a cluster of adjacent pixels that share similar visual properties, i.e. color, texture, etc. The proposed superpixel voting procedure is summarized as follows, and the schematic diagram is shown in Fig. 6:

1. Choosing a frame with a few objects and smoothing the frame using the median filter.
2. Segmenting the image into superpixels by simple linear iterative cluster (SLIC) algorithm, proposed by Achanta et al. (2012). The approximate number of segmentations is the hyperparameter of SLIC.
3. Duplicating the superpixel segmentation to all frames.
4. Computation of structural similarity (SSIM) for each pair of corresponding superpixels between the current frame and every other. If the SSIM larger than 0.65, the votes of this superpixel increase by one.
5. Reconstruction of the result: the result superpixel is the median of the superpixels with the most votes; the image is reconstructed by jointing all the result superpixels. Because of the commutativity of SSIM, there must be more than one superpixel having the most votes.

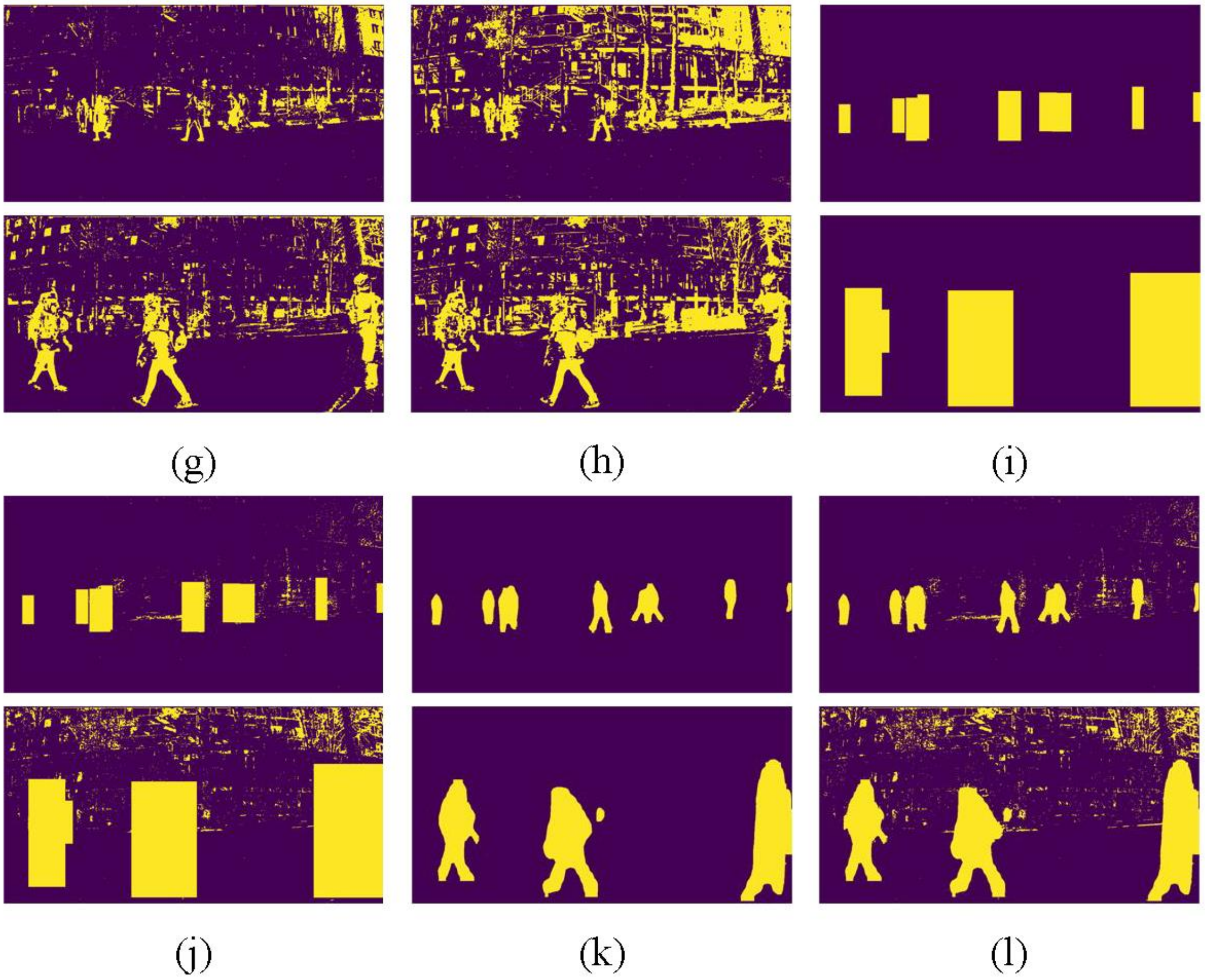

Fig. 8. Examples of obtained masks for two picture following described methods: (g) KNN (training images initialization), (h) KNN (gamma initialization), (i) YOLO, (j) YOLO with shadow detection, (k) Mask R-CNN, and (l) Mask R-CNN with shadow detection (© Zheng Zhang).

All the approaches above will be evaluated in the section of experiments. Figure 7 and 8 gives examples of masks obtained according to these different methods.

Results and discussion

This section demonstrates the experimental results of the moving object removal approaches and the panorama constructed from the pure background images. Different approaches are evaluated quantificationally and analyzed.

The experimental dataset is captured by the fish-eye lens pose by pose under the control of the robotic head, using multiple passes, producing an image sequence for each camera pose. Each acquisition pass consists of several camera poses, and multiple passes are executed at the same capturing position, i.e., there is an image sequence for each camera pose in the dataset. To extract background information from these image sequences, the first requirement for the acquisition of data is that there are enough photos for each spatial point to extract the background from. The second requirement is that each pair of adjacent photos has an overlap so that the photos can be stitched to a panorama.

To realize our testing datasets, we used a Panocatcher Maestro 4HD heavy-duty, full-size, dual-axis robotic head with a Zenitar 16mm f/2.8 fish-eye lens mounted on a Nikon D810. The robotic panoramic head offers the possibility to make each time the same shots at almost the same determined rotation angle. The state-of-the-art stitching pipeline (Fig. 3) is used to obtain the panoramas.

The ground truth should be a pure background without any foreground objects. If the dataset is captured in a place with high traffic, it will tough to find an opportunity that no one passes by. A solution to address this problem is proposed: taking out the photos that have small amounts of moving objects as a test set, the others who have plenty of moving foreground are put in the training set. The ground truth is the median image of photos in the test set, which is an optimal background. Some pedestrians are added manually to part of the training images to create critical scenarios.

The used metrics are root-mean-squared error (RMSE) and structural similarity (SSIM). RMSE can measure the average error. However, sometimes human eye judgment may be different from numerical error, so the structural similarity (SSIM) is used to consider the luminance, contrast, and structure comprehensively. For the RGB image, the MSE of each channel is summed and divided by three.

The quantitative evaluation is shown in Table 2. The result of the median approach has lots of ghosts, and the ghost problem arises in the areas where have moving objects frequently. Most of the foreground-mask-based approaches perform better than the median. The performance of foreground-mask-based approaches depends on the quality of the masks, so the two CNN-based approaches outperform others with the most precise masks. The masks of YOLO are composed of filled bounding boxes whose edges may leave some traces on the result. The masks of Mask R-CNN are in line with the shape of the objects, which will not waste background information and will generate a cleaner result.

Table 2. Quantitative evaluation, the bold values are the best in the metric.

Algorithm	**RMSE**	**SSIM**
Median	6.5862	0.9127
Three-frame differencing	7.0378	0.8982
MOG2 (median init)	7.1981	0.9051
MOG2 (training images init)	7.2795	0.9064
MOG2 (gamma adjusted medians init)	6.4634	0.9093
KNN (median init)	7.4758	0.8891
KNN (training images init)	7.6481	0.8962
KNN (gamma adjusted medians init)	8.0399	0.8948
YOLO (without shadow detection)	6.0042	0.9179
YOLO (with shadow detection)	6.1213	0.9101
Mask R-CNN (without shadow detection)	***5.9205***	***0.9182***
Mask R-CNN (with shadow detection)	*6.0105*	*0.9115*
Superpixel voting (1000 segments)	6.8101	0.9091
Superpixel voting (2000 segments)	6.7598	0.9076
Superpixel voting (3000 segments)	7.0700	0.9049

The defects of the superpixel voting approach are in the form of patches, which are in the shape of the superpixel segment. The number of segments determines the performance: too few segments will increase the error; too many segments will be time costing. Another advantage of YOLO and

Mask R-CNN is that they do not require multiple frames to detect moving objects, allowing the problems detection in parallel with acquisition. In conclusion, the Mask R-CNN approach is the most accurate, and the shadow can be detected by MOG2.

Conclusions

The main contributions of this work include: (1) data acquisition method and semi-synthetic dataset generation method for evaluation; (2) the initialization approaches to apply the background modeling approaches on a sequence of few pictures; (3) proposing, implementation and evaluation of the moving foreground removal approaches. The image sequence is aligned and split into a test set and training set, and the images in the test set are used to generate ground truth. Synthetic foreground objects are added to the training set for critical evaluation. All the proposed moving object removal approaches are evaluated and analyzed. Best results are obtained combining both approaches: Mask R-CNN for object detection and MOG2 for shadow detection

The experimental results (Fig. 1 and Fig. 10) show that the ghosting artifacts and the foreground objects have been effectively removed using the approach in this paper.

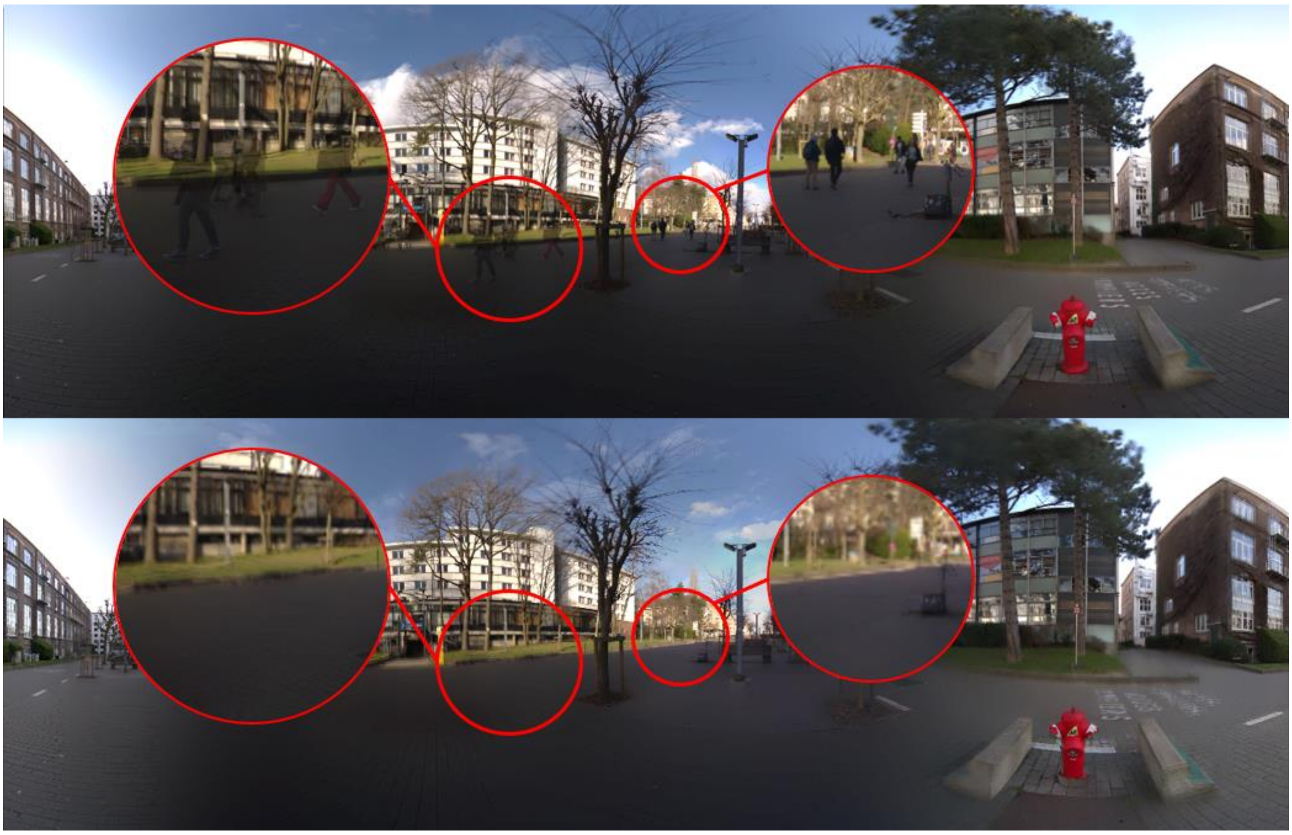

Fig. 10. Comparison of panorama with traditional (top) and proposed method (bottom) (© Zhang Zheng).

Funding

The project did not receive external funding.

Conflict of Interests Disclosure

No conflicts of interests have been declared by the authors.

Author Contributions

Conceptualization: Arnaud Schenkel
Software: Zheng Zhang
Supervision: Arnaud Schenkel, Olivier Debeir
Validation: Zheng Zhang
Writing – original draft: Zheng Zhang
Writing – review & editing: Arnaud Schenkel

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). 'SLIC Superpixels Compared to State-of-the-Art Superpixel Methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), pp. 2274–2282.

Chen, Y., Chen, Y., and Wang, G. (2019). 'Bundle Adjustment Revisited', [online] Available at: arxiv:1912.03858 (Accessed: 15 January 2021).

El Abbadi, N., Al Hassani, S., and Abdulkhaleq, A., (2021). 'A Review Over Panoramic Image Stitching Techniques', *2nd International Virtual Conference on Pure Science (2IVCPS 2021), Journal of Physics: Conference Series*, Volume 1999.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). 'Mask R-CNN. 2017', *IEEE International Conference on Computer Vision* (ICCV 2017).

Herrmann, C., Wang, C., Bowen, R.S., Keyder, E., and Zabih, R. (2018). 'Object-Centered Image Stitching', in *ECCV 2018. Lecture Notes in Computer Science*, vol. 11207, pp. 846–861.

Kameda, Y. and Minoh, M. (1996). 'A human motion estimation method using 3-successive video frames', *International conference on virtual systems and multimedia*, pp. 135–140.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). 'You Only Look Once: Unified, Real-Time Object Detection', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Su, T. (2018). 'A Superpixel-Based Voting Scheme for Removing Moving Objects in Photos', *Proceedings of the 2018 2nd International Conference on Advances in Energy, Environment and Chemical Science (AEECS 2018)*.

Szeliski, R. and Shum, H. (1997). 'Creating Full View Panoramic Image Mosaics and Environment Maps', *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*.

Szeliski, R. (2007). 'Image Alignment and Stitching: A Tutorial', *Foundations and Trends® in Computer Graphics and Vision*, 2(1), pp. 1–104.

Zivkovic, Z. and van der Heijden, F. (2006). 'Efficient adaptive density estimation per image pixel for the task of background subtraction', *Pattern Recognition Letters*, 27(7), pp. 773–780.

Zivkovic, Z. (2004). 'Improved adaptive Gaussian mixture model for background subtraction', *Proceedings of the 17th International Conference on Pattern Recognition*, 2004. ICPR 2004.