The method to annotate buildings with images from CCTV cameras with an intersecting field of view

N V Nazarenko[1,2], D E Bekasov[1]

[1]Bauman Moscow State Technical University, 105005 Baumanskaya 2th st., 5, Moscow, Russian Federation

[2]E-mail: nazarenkonv@student.bmstu.ru

Abstract. Video surveillance and tracking systems help to increase the safety of human activity in a given area and improve detecting the geo-location of an observer (a swarm of drones, autopilot) in modern cities. This article proposes the method for determining the full or partial address (annotation) of buildings by using images from several CCTV cameras with an intersecting field of view. As input parameters, the method takes several images from surveillance cameras, their geographical coordinates and directions of the optical axes and the vector map of the area where the cameras are located. The outputs are annotations of buildings for the initial images. The proposed method allows solving the problem of georeferencing the video analytics system and improving the work of the surveillance and object tracking systems. The resulting annotations of buildings provide an opportunity to determine more accurately the situations with overlapping of the tracked objects with buildings and predict the probable paths of the objects. A lost object on one camera can be found by using annotations from other cameras, which field of view is not overlapped by this building. Besides, the proposed ap0070roach is sustainable to full or partial overlapping of buildings and allows solving to solve the problem in the absence of a detailed height map of the terrain and buildings. Also, the article researches the restrictions of the method. It is necessary to find an angle between two cameras, at which the best result can be achieved. That is, the method should return the correct addresses of all buildings visible in the image.

1. Introduction

Video analytics is used in such spheres of modern life activity as retail and service, logistics, and security. In the field of security, video analytics can be used to analyze the video image of a limited area automatically [1]. The main functions are:

- Counting objects number in the control zone;
- Objects identification located in the control zone, according to specific criteria (personnel detection by uniform, etc.);
- Events detection in the specified perimeter (motion detection, objects absence in the perimeter, etc.);
- Tracking the movement of the object;
- Objects geolocation in the given perimeter.

The modern video analytics systems are required to automate the geolocation function. For example, most systems use a manual setup of camera geolocation and its optical axes in the task of georeferencing. Geolocation is required in the tracking task to improve the quality of object tracking. This task can be divided as follows:

- Object paths construction;

- Classification of situations with object loss (overlapping by an obstacle, algorithm error, etc.).

This article proposes a method to annotate buildings using CCTV cameras with an intersecting field of view and researches the restrictions of the method. This method allows classifying situations with objects loss and, in some cases, avoiding it. Besides, it can be used to detect which buildings are in the camera field of view.

2. Related Work

The existing systems available in open sources allow solving the problem only partially:

- VisualSFM [2], Bundler [3], and other similar systems based on the SfM (Structure from Motion) method are systems for 3D reconstruction [4] as a point cloud. There are two types of reconstruction: sparse and dense. Such systems take a set of images, their elements or matches as input data and create a 3D reconstruction of the camera and scene geometry as a result. Such systems are successfully used to create the first version of a scene point cloud further texturing it.
- Mask R-CNN is a neural network based on a convolutional neural network. Mask R-CNN segments objects, even if there are several instances, they have different sizes and partially overlap [5].

As already mentioned above, the existing solutions solve the indicated problem only partially. They either detect and segment the building in the image or determine the relative depth of a particular image point (pixel). Thus, none of the systems determines which buildings are being observed in the image.

3. The Method to Annotate Buildings with Images from CCTV Cameras with an Intersecting Field of View

Two cameras with the intersecting field of view can be used to solve the task of annotating buildings. It does not matter which camera to use (CCTV, mobile, etc.) for the method to work. The main limitation imposed on the camera is the high resolution of frames obtained from the video stream of the camera. The higher the resolution, the more characteristic points can be highlighted. Therefore, the whole method will work more accurately. Then images taken at any time from these two cameras are processed by this method and annotations of the building are returned.

The following scheme can be proposed for the buildings annotation method based on the above-mentioned partial solutions, which consists of the next steps:

1. Building a point cloud of the area by images from CCTV cameras to obtain the relative depth of each image point;
2. Filtering a point cloud to clean the cloud from noise and points do not belong to buildings. These points may belong to cars or people;
3. Clustering of individual buildings from a point cloud to determine cloud points association to specific real buildings;
4. Comparison of clusters and area vector map where the cameras are located to determine the annotations of buildings.

The method takes area images from the video surveillance cameras, these cameras coordinate and the directions of their optical axes, and a vector map as input data. The method result is the annotations of buildings.

3.1. Point cloud

Initially, it is necessary to determine the relative depth of each building seen by the cameras. To do this, it needs to find the relative depth of each point in the image, or other words, to build a point cloud [6] of the area observed by cameras.

The set of photogrammetric methods SfM and MVS (Multi-view Stereo [7]) was chosen to build a point cloud. These methods do not require camera mobility or any additional sensors (lidars, IR sensors). Moreover, they are not affected by the external factors interfering with the method (scattered light, reflective surfaces, etc.). At the same time, there is a requirement for the presence of two or more cameras with an intersecting field of view [8]. The result of this step is a dense point cloud, built on the input images.

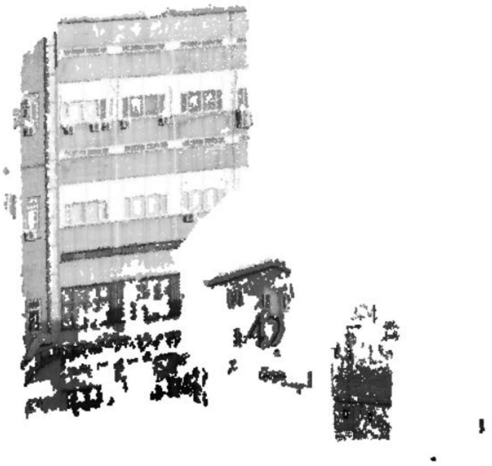

Figure 1. An example of a dense area reconstruction from two CCTV cameras with an intersecting field of view.

The SfM method in this set is used to obtain such camera parameters as its location in the point cloud and its projection matrix. This matrix with 3x4 dimension contains a camera rotation angle relative to the origin, focal length, and distortion. Then the MVS method uses the obtained parameters to reconstruct a dense point cloud (figure 1).

3.2. Point cloud filtration

The resulting point cloud must be corrected by removing the points formed by the error of cloud reconstruction methods (“outliers”) and points belonging not to buildings, but any other objects from this cloud. For example, to people or cars.

In order to keep in the cloud the points belonging only to buildings, it was initially proposed to use the filtering cluster method. The idea is based on the fact, that clusters, describing buildings, are much larger than clusters, belonging to cars or people. However, it is impossible to determine the exact cluster size that can be considered large enough for a building. It is also quite possible when reconstructing a point cloud that one of the buildings consists of much smaller points number than the others. In this case, the points that need to be saved will still be deleted.

Another option of the filtering method is the edge detectors usage. However, since the outlines of buildings are very different, it is difficult to determine whether this edge belongs to any building.

The most suitable variant for filtering a point cloud is the neural networks usage. Neural networks can find buildings on the images and segment them. So it is possible to save the points in the cloud belonging to the buildings. In addition, this method can also filter out “outliers” resulting from errors in the cloud reconstruction methods.

Mask R-CNN was chosen for the filtering algorithm implementation. It is a neural network architecture based on a convolutional neural network [9]. It can segment objects on images (determines whether an image pixel belongs to this object or not). Mask R-CNN takes any RGB image as input.

The proposed filtering algorithm consists of two main parts. Firstly, it segments objects in the input images from surveillance cameras using the Mask R-CNN. Secondly, it removes points not belonging to buildings from the dense cloud obtained in the first step of the annotation method. As a result, the filtering algorithm returns a dense cloud containing points that belong only to buildings.

To remove extra points from a dense cloud, we can use the projection matrix *P* of the camera from the SfM algorithm output. This matrix can be used to accurately determine the point location on the image caught by this camera using its coordinates in the cloud (equation (1)):

$$d\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = (P) * \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}, \tag{1}$$

where *u, v* denotes coordinates of the pixel in the image; *d* is the point depth with respect to the current camera; *P* denotes a 3x4 projection matrix of the camera; *x, y, z* denotes coordinates of the point in the cloud.

3.3. Buildings clusterisation from point cloud

Filtered point cloud needs to be clustered into separate buildings. This step is necessary to determine the affiliation of points to real buildings.

The k-means [10] algorithm was chosen for the clustering realization. Moreover, the neural network removes noises from the cloud that could affect the quality of the k-means algorithm. As a metric, we use the Euclidean distance [11] in three-dimensional space (equation (2)):

$$d_{ij} = \left[\sum_{i=1}^{3} (p_{il} - p_{jl})^2 \right]^{1/2}, \tag{2}$$

where p_i is the i-th point in the cloud, and *l* is the component (coordinate) x, y or z of this point.

3.4. Matching of clusters and map

The last step of the method to annotate buildings is necessary for matching clusters from the previous step with the original buildings on the map. No algorithm was found in open sources that solves the similar problem. Therefore the proprietary solution was proposed.

Initially, we know the real geographical coordinates of the cameras and their optical axes direction. From the SfM output, we also know the coordinates of the cameras in the point cloud space. So it is possible to obtain the transition matrix *MT* from the geographical coordinates to the point cloud coordinates and the scaling factor of the coordinate system. It suffices with the matrix *MT* to take the central or middle point k_i of each cluster and apply the transformation into the geographic coordinates to this point, given the calculated scaling factor. The resulting coordinates using the reverse geocoding by the vector map determines the current building address. The pixel coordinates of the annotations are then calculated using the projection matrix obtained earlier from the SfM method.

As a geographic coordinate system, this algorithm can use the ECEF (Earth-Centered, Earth-Fixed [12]). The coordinates of the point cloud can be represented by the rotated ENU system (East, North, Up [12]). Then, using the available input data, it will not be difficult to calculate a transition matrix from one system to another.

4. Restrictions of the Method

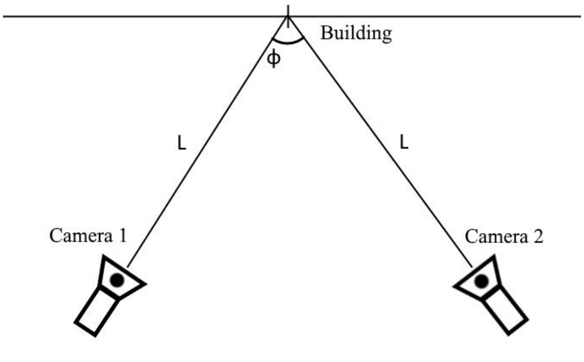

Figure 2. The position of the cameras during the test environment creating.

One of the most critical parameters of the system configuration, which affects the method applicability, is the angle φ between the cameras (figure 2). Therefore, it is necessary to find such an angle, so the result obtained by the method to annotate buildings is the best. The sample for the experiment consists of 120 images of different scenes and buildings. The following characteristics are used as criteria for evaluating the performance of the method:

- The number of points in the constructed cloud. The more points cloud will contain, the more buildings we can identify and annotate. Therefore, the result will be better and more accurate;
- The correctness of the received addresses of buildings.

A generalized graph of the experiment results is presented in figure 3. It can be seen that the highest number of points is reached at an angle of 30 degrees. The angles of 0 and 90 degrees show the worst results. At these angles, the number of constructed points is zero. Therefore, the address of the building in the image cannot be determined at all. In all other cases, the addresses of buildings were detected correctly. Hence the method to annotate buildings worked correctly. Thus, the proposed method can work correctly at angles between two cameras from 15 to 60 degrees. However, the best result achieved at the angle of 30 degrees. These restrictions make it possible to apply this method in real-life tasks when cameras with an intersecting field of view are located on different sides of the same building or on neighbouring buildings placed close to each other.

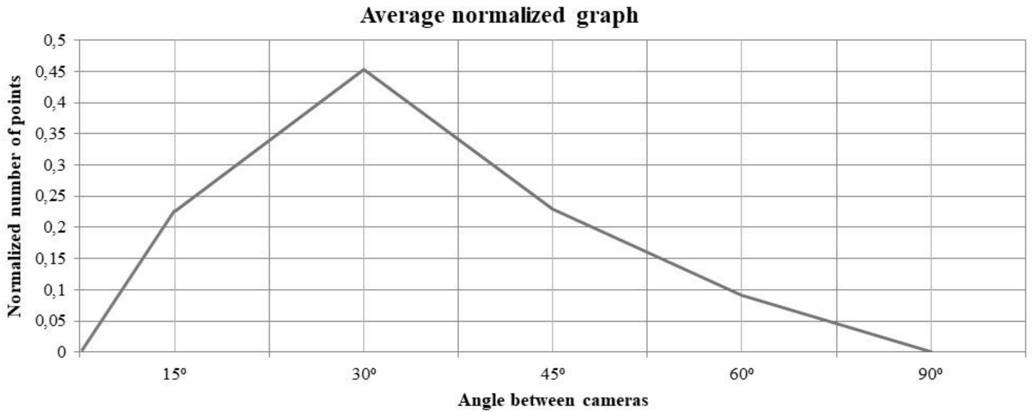

Figure 3. The average normalized graph of the number of points versus the rotation angle.

The ratio of correctly determined addresses to incorrect (accuracy of the method) on the images used above was 0.83. As already noted, mostly incorrect addresses are determined in images made at angles of 0 and 90 degrees.

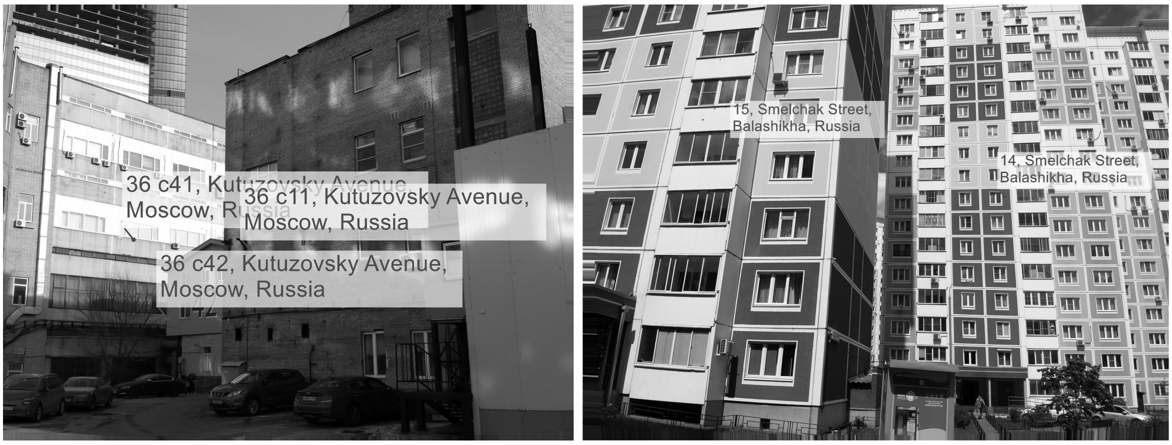
36 c41, Kutuzovsky Avenue, Moscow, Russia
36 c11, Kutuzovsky Avenue, Moscow, Russia
36 c42, Kutuzovsky Avenue, Moscow, Russia
15, Smelchak Street, Balashikha, Russia
14, Smelchak Street, Balashikha, Russia

Figure 4. Example results of the method to annotate buildings.

Examples of the proposed method to annotate buildings are shown in figure 4. Annotation attaches to each building on each initial image and consists of the number, street, city, and country where the building is located. It can be accurately determined when and which building overlaps the tracked object by using these annotations. This information can be further used to detect an overlapped object on the camera with the field of view that is not blocked by this building. Thus, cameras quickly locate the object, and moreover, it is always possible to predict its probable path.

5. Conclusion

This article has proposed the method to annotate buildings with images from CCTV cameras with an intersecting field of view and reviewed all its steps. Also, this article has researched the restrictions of the proposed method at different angles between two cameras to establish the range of applicability. During this study, the accuracy of the method was found.

As a further development of this method, it is possible to modify the neural network so that it would work directly with point clouds, and not with input images, as it has been done in the current implementation. This modification will increase the overall speed of the method.

References

[1] Taranyan A R, Devyatkov V V, Alfimtsev A N 2018 Selective Covariance-based Human Localization, Classification and Tracking in Video Streams from Multiple Cameras *Proc. of the 11th Int. Joint Conf. on Biomedical Engineering Systems and Technol. (BIOINFORMATICS* vol 4*)* (Funchal: SciTePress) pp 81–8 http://dx.doi.org/10.5220/0006538100810088

[2] Wu C 2019 *VisualSFM: A Visual Structure from Motion System* http://ccwu.me/vsfm

[3] Snavely N 2019 *Bundler: Structure from Motion (SfM) for Unordered Image Collections* http://www.cs.cornell.edu/~snavely/bundler

[4] Ostroukh A, Vakhrushev O, Maikov K, Kolbasin A 2019 Combined identification method by reconstruction and analysis of face 3D structure *ARPN J. of Engineering and Appl. Sci.* **14** pp 1385-1388

[5] He K, Gkioxari G, Dollar P, Girshick R 2017 Mask R-CNN *IEEE Int. Conf. on Computer Vision (ICCV)* (Venice: IEEE) pp 2980-2988 http://dx.doi.org/10.1109/iccv.2017.322

[6] Kozov A V, Volosatova T M, Vukolov A Y 2018 Structural Obstacle Recognition Method and Its Application in Elevated Terrain Objects Search *Int. Russian Automation Conf. (RusAuto Con)* (Sochi: IEEE) pp 1-5 http://dx.doi.org/10.1109/ rusautocon.2018.8501765

[7] Furukawa Y, Hernández C 2015 *Multi-View Stereo: A Tutorial (*Now Publishers) p 166

[8] Kuznetsov A O, Gorevoy A V, Machikhin A S 2019 Image rectification for prism-based stereoscopic optical systems *Computer Vision and Image Understanding* **182** pp 30-7

[9] Anishchenko L 2018 Machine learning in video surveillance for fall detection *Ural Symp. on Biomedical Engineering, Radioelectronics and Information Technol. (USBEREIT)* (Yekaterinburg: IEEE) pp 99-102

[10] Piech C 2019 *K Means* https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

[11] Barret P 2019 *Euclidian Distance* https://www.pbarrett.net/techpapers/euclid.pdf

[12] Leick A, Rapoport L, Tatarnikov D 2015 *GPS Satellite Surveying* (John Wiley & Sons, Inc) p 807