Article

Extracting Objects' Spatial–Temporal Information Based on Surveillance Videos and the Digital Surface Model

Shijing Han ^{1,2}, Xiaorui Dong ¹ , Xiangyang Hao ^{1,*} and Shufeng Miao ³

¹ Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China; shijinghan@nnnu.edu.cn (S.H.); dongxiaorui@slcupc.edu.cn (X.D.)

² School of Natural Resources and Surveying, Nanning Normal University, Nanning 530001, China

³ Wuhan Kedao Geographical Information Engineering Co., Ltd., Wuhan 430081, China; shufengmiao@geniuses.com.cn

* Correspondence: xiangyanghao2004@163.com

Abstract: Surveillance systems focus on the image itself, mainly from the perspective of computer vision, which lacks integration with geographic information. It is difficult to obtain the location, size, and other spatial information of moving objects from surveillance systems, which lack any ability to couple with the geographical environment. To overcome such limitations, we propose a fusion framework of 3D geographic information and moving objects in surveillance video, which provides ideas for related research. We propose a general framework that can extract objects' spatial–temporal information and visualize object trajectories in a 3D model. The framework does not rely on specific algorithms for determining the camera model, object extraction, or the mapping model. In our experiment, we used the Zhang Zhengyou calibration method and the EPNP method to determine the camera model, YOLOv5 and deep SORT to extract objects from a video, and an imaging ray intersection with the digital surface model to locate objects in the 3D geographical scene. The experimental results show that when the bounding box can thoroughly outline the entire object, the maximum error and root mean square error of the planar position are within 31 cm and 10 cm, respectively, and within 10 cm and 3 cm, respectively, in elevation. The errors of the average width and height of moving objects are within 5 cm and 2 cm, respectively, which is consistent with reality. To our knowledge, we first proposed the general fusion framework. This paper offers a solution to integrate 3D geographic information and surveillance video, which will not only provide a spatial perspective for intelligent video analysis, but also provide a new approach for the multi-dimensional expression of geographic information, object statistics, and object measurement.

Keywords: fusion; surveillance video; 3D geographic information; mapping; moving object; DSM

Citation: Han, S.; Dong, X.; Hao, X.; Miao, S. Extracting Objects' Spatial–Temporal Information Based on Surveillance Videos and the Digital Surface Model. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 103. <https://doi.org/10.3390/ijgi11020103>

Academic Editor: Wolfgang Kainz

Received: 17 November 2021

Accepted: 30 January 2022

Published: 2 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Real-time video surveillance plays an increasingly important role in crime prevention, traffic control, environmental monitoring, terrorist threats, and city management [1]. Video surveillance is an effective tool for real-time monitoring 24 h a day [2], and cameras are the visual organs of smart cities [3]. These cameras collect a large amount of video data, and it is essential to extract useful information from the massive data. Computer vision replaces human eyes for visual recognition, tracking, and the measurement of video objects [4]. In recent years, object detection and tracking have become the research hotspot and frontier in the field of computer vision. However, due to the influence of many factors, such as occlusion, motion blur, illumination change, scale change, etc., missed detection and wrong matching still occur in multi-object tracking. In addition, video surveillance systems connect each camera to a corresponding monitor in the control center, which cannot reflect the spatial relationship between different cameras in the geographical scene. Therefore, video observers must understand the monitoring area to mentally map the images on each monitor to the corresponding region in the real world [5]. When there are enough

cameras, there is no doubt that it is a considerable challenge. At the same time, with the rapid development of surveying and mapping technology, people put forward higher requirements for the expression of spatial geographical entities. More practical and creative elements such as street view, remote sensing images, and 3D models are integrated into map-making, enhancing the visual expression of a map. However, the current geographic information is still mainly represented by static objects. The basic geographic information data need to be measured in advance and stored in the database, which is characterized by high accuracy, a unified coordinate framework and poor real-time performance.

Based on the above analysis, video data are intuitive, informative, and highly redundant, but they can record the movement of moving objects, such as vehicles and pedestrians in real-time. However, geographic information can only express the static geographic space in the surveillance video. How to describe the moving objects in a surveillance video in a static geographical framework is a topic worthy of study. GIS was suggested as a general frame of reference to which all cameras could be mapped [6]. The framework not only provides a unified spatial reference, but also provides rich semantic information, which facilitates multi-camera collaboration and object tracking. The integration of static geographic information and dynamic video enables the expression of moving objects in a surveillance video under a high-precision and unified framework. Not only does the integration of the two provide data support and a guarantee for more industries (fields), larger scales (space and time) and more dimensions of deep learning research, but these big data, with a unified time baseline, geographic reference frame, and intrinsic logical relationship, will undoubtedly have a profound impact on the development of surveying and mapping.

Some scholars and industry practitioners at home and abroad have conducted research on this topic, but there are still some problems. In terms of video geo-spatialization, many research studies are based on the assumption of a planar ground in geographic space. The mapping relationship between images and the real world is established through a homography matrix [7,8], that is, from a 2D image to 2D space. The homography matrix-based method is not suitable for large-scale scenes or scenes with complex terrain. In the aspect of fusion, the research is relatively scattered, as it is used for multi-camera object tracking [9], path searching [10], video fragment data management [11], video synopses [12], crowd counting [13], etc. The current real world is in rapid development and change, and how to express and recognize it thoroughly is very important. The real 3-dimensional city model (3DCM) has become a hotspot of current surveying and mapping research. The 3D geographical scene is the development demand of geographical information at present. It is clear that the integration of moving objects and 2D static geographical scenes cannot meet this demand. Exploring the effective fusion of moving objects in a surveillance video and a 3D static geographical scene has become a problem to be solved at present.

In this work, our research mainly includes three parts. Firstly, we detect and track objects based on deep learning to obtain the object information in the image. Second, we rely on the high-precision digital surface model (DSM) to establish the mapping model between the surveillance video and the 3D geographic space, that is, from 2D images to 3D space. Thirdly, based on the mapping model, the spatial-temporal information of moving objects is visualized in a 3D model. Experiments show that the proposed methods achieve good results. The methods proposed are helpful for users to understand the activities of moving objects in a geographical scene quickly and efficiently, which allows users to feel them in person.

The remainder of this paper is organized as follows: Section 2 introduces related works. Section 3 presents a fusion framework of 3D geographic information and moving objects in surveillance videos. Section 4 illustrates the principles and methods adopted in this paper from four aspects: the camera imaging model, the ray intersection with the DSM, pedestrian detection and tracking, and the acquisition of objects' spatial-temporal information. Section 5 conducts and analyzes the experiment. Section 6 presents trajectories of moving objects in the 3D model. Section 7 summarizes and concludes the study.

2. Related Work

There are some advantages to surveillance video data, such as richness, intuition, and the timeliness of the information. However, there are also several challenges, such as the massive volume of data, the scarcity of geographic information, and the sparseness of high-value information. It is necessary to extract moving objects from surveillance videos and geo-spatialize video images. In this section, we introduce related works on four aspects: camera calibration, video geo-spatialization, object detection and tracking, and the integration of surveillance videos and geographic information.

2.1. Camera Calibration

Camera calibration is a fundamental topic in the field of computer vision and is essential in many applications, such as video surveillance, three-dimensional reconstruction, robot navigation, etc. By doing this, the intrinsic and extrinsic parameters of the camera can be obtained. Intrinsic parameters include the focal length, principal point, skew coefficients, and distortion coefficients, which are the camera's inherent characteristics. Zhang [14] proposed a flexible technique to calibrate a camera by observing a planar pattern shown at a few different orientations, which is easy to use and has high precision. Hartley's method [15] was carried out by analyzing point matches between at least three images taken from the same spatial point with different camera directions, which was based on the pure rotation of the camera without knowledge of its orientations. Triggs [16] first introduced the absolute quadric into a self-calibration field, a method which needs at least three images taken by a moving camera with fixed but unknown intrinsic parameters. The intrinsic parameter calibration of the camera has been well described in the literature. Extrinsic parameters determine the camera's position and orientation in the world. They can be obtained by extrinsic devices (GPS and IMU) or the Perspective-n-Point (PnP) [17] problem in computer vision. Some scholars have studied the PnP problem. Lepetit [18] proposed a noniterative solution from n 3D-to-2D point correspondences, which was applicable for all $n \geq 4$ and handled both planar and non-planar configurations properly. Li's solution [19] for the PnP problem was also non-iterative and could robustly retrieve the optimum. When there were no redundant reference points, its results were better than the iterative algorithm.

2.2. Video Geo-Spatialization

Video geo-spatialization studies the mapping relationship between image points and spatial points. There are two main methods of video geo-spatialization: the method based on the homography matrix [7,8] and the method based on the intersection between the imaging ray and terrain model [20]. The former assumes that the ground in space is flat, and generally assigns its elevation to 0. The homography matrix can be determined by four or more image coordinates corresponding to the world coordinates. Although the calculation is small, it is not suitable for large or complex topographic scenes. The latter requires the construction of imaging rays and a traversal search on the terrain model. Consequently, this method is not affected by topography. Nonetheless, it requires a large amount of computation and a high-precision terrain model. In recent years, other mapping methods have emerged. Milosavljević [21] adopted a reverse process by back projecting the position-determined objects onto the video image. In 2017, Milosavljević [22] estimated surveillance georeferences by pairing the image coordinates with their 3D geographic locations, which could be used for georeferencing fixed and PTZ surveillance cameras.

2.3. Object Detection and Tracking

Object tracking is a field of computer vision that aims to maintain objects' identities. Tracking-by-detection is a widely used tracking framework, in which objects are first detected and then linked into trajectories. Object detection requires solving two problems: where the object is and what the object is. In 2014, Girshick [23] first proposed R-CNN, which can be used in object detection models, and started the upsurge in deep

learning-based object detection. Although Fast R-CNN [24] was committed to reducing the running time of an object detection network, it still faced a bottleneck of region proposal computation. To solve the problem, Ren [25] introduced a region proposal network (RPN) for region proposal generation, which shared full-image convolutional features with Fast R-CNN. He [26] proposed Mask R-CNN, which added only a small overhead to Faster R-CNN, and could detect objects efficiently and obtained a high-quality segmentation mask simultaneously. The detectors above are all the two-stage models of object detection frameworks. Based on regression, Redmon [27] first presented the You Only Look Once (YOLO) object detection model. Because it only has one training network, YOLO has a vast advantage in terms of execution speed. Given the deficiency of YOLO in small object detection, the SSD [28] model adopted the regression idea used in YOLO and referred to the anchor mechanism proposed in Faster R-CNN. YOLOv2, YOLOv3, YOLOv4, and SSD pertain to single-stage models. Object tracking is used to locate the place of the object and to record the trajectory and parameters of the object of interest. Matching tracklets to generate a complete global trajectory is a problem. Huang [29] first proposed the hierarchical association algorithm, which linked detection responses into tracklets, and then these highly fragmented tracks were further correlated at each level of the hierarchy to generate the final long trajectory. He [30] proposed the restricted non-negative matrix factorization (RNMF) algorithm to solve the tracklet matching problem by reducing tracking errors in tracklets. Xu [31] proposed an intuitional but easy to implement method called feature group, which reduced the decline in accuracy due to occlusion. Wang [32] allowed object detection and appearance embedding to be learned in a shared model, which was the first (near) real-time MOT system, with a running speed of 22 to 40 FPS and a high tracking accuracy. When the topological relation of cameras is unknown, the multi-camera tracking problem can be abstracted into a person re-identification (re-ID) problem. Ristani [33] used weighted triplet loss and a hard-identity mining technique to yield appearance features, which performed well in object detecting and re-ID. Zhang [34] used Faster R-CNN to detect objects and a person re-identification model to extract the appearance features first, and then merged the trajectories by hierarchical clustering. Tagore [35] proposed an efficient hierarchical re-identification approach, first through color histogram and then through deep feature-based comparison, which was evaluated on six datasets.

2.4. The Integration of Surveillance Videos and Geographic Information

Katkere [36] integrated GIS and video for the first time by using multiple video data streams to create immersive virtual environments. Takehara [37] integrated fixed cameras and real-time captured data streams, which could build three-dimensional views and present people's movement in space in an understandable way. Zhang [38] assumed that the rotation of the camera was zero and then calculated the transformation matrix to realize the integration of 2D GIS and video surveillance. Yang [39] adopted multi-view surveillance and rendered steerable 3D views of tracked objects onto a reconstructed 3D site representation, which provided users with viewpoints as the event occurred. Xie [8] discussed the integration of GIS and moving objects in surveillance video and proposed an integrating model in which video data were acquired by a single camera. In 2019, Xie [7] discussed the integration of multi-camera video moving objects (MCVO) and GIS. Application scenarios of both papers are restricted to a situation in which the video area is flat.

3. Framework of 3D Geographic Information and Moving Objects

Video information can be collected by cameras from different geographic locations, which is a sequence of images (frames). If security operators are not familiar with the monitoring area, they cannot obtain the location of the objects from these images. The integration of a video and GIS can unify the objects that they are interested in. By doing this, the spatial-temporal information of moving objects can be displayed in a 3D geographic model, which can help to implement the measurement of objects and enhance the

visual multi-dimensional representation of GIS. The fusion framework of 3D geographic information and moving objects in surveillance video is shown in Figure 1.

Figure 1. Fusion framework of 3D geographic information and moving objects in surveillance video.

In our framework, we divide the fusion of 3D geographic information and moving objects in surveillance video into several parts:

- Camera calibration. Camera calibration aims to obtain intrinsic and extrinsic parameters which can determine the camera model.
- The extraction of moving objects. Object detectors and trackers can extract the information of moving objects in a video which is about spatial-temporal positions in an image. These positions are not very accurate and contain errors. They need distortion correction.
- Mapping model. In the process of camera imaging, depth information is lost. To recover the position of the object in space, we need terrain information obtained by a 3D geographic model. Given a calibrated camera and an image pixel, an imaging ray $(X_0, Y_0, Z_0) + k(U, V, W)$ is constructed. The imaging ray intersects the terrain at a point where the object is located.
- Spatial-temporal information in 3D space. In addition to the 3D position information, the width and height of the object can also be obtained by geometric calculation.
- Visualization. Present moving objects in a 3D geographic scene.

In summary, camera calibration and the extraction of moving objects is the basis of fusion; the mapping model is a bridge from image space to the real world; spatial-temporal information in 3D space is the critical data of fusion; and visualization is the result of fusion. Obviously, the fusion of 3D geographic information and moving objects in surveillance video aims to provide a unified framework and make the advantages of the two complementary.

4. Principles and Methods

Section 3 has shown the fusion framework and presented a general description. The framework does not rely on specific algorithms for the camera model, object extraction, or the mapping model. In Section 4, we will elaborate on the principles and methods of these core issues in this paper.

4.1. Camera Imaging Model

A camera is a mapping between the 3D world (object space) and a 2D image. In other words, we can map points in space with three-dimensional coordinates to points on the image plane under the camera model. The most specialized and simplest camera model is the pinhole camera. The imaging model is listed in Equation (1):

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & s & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_{1 \times 3} & \mathbf{I} \end{bmatrix} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix} = \mathbf{C} \begin{bmatrix} X_W \\ Y_W \\ Z_W \\ 1 \end{bmatrix}, \quad (1)$$

where λ in Equation (1) is a scale factor; $\begin{bmatrix} u & v \end{bmatrix}^T$ is a point on the image plane; $\begin{bmatrix} u_0 & v_0 \end{bmatrix}^T$ are the coordinates of the principal point O ; f_x and f_y represent the focal lengths of the camera in terms of pixel dimensions in the u and v direction, respectively; the parameter s is the skew parameter; and $\begin{bmatrix} X_W & Y_W & Z_W \end{bmatrix}^T$ is a point in space. The parameters contained in $\mathbf{K}$ are called the intrinsic parameters. The parameters contained in $[\mathbf{R}|\mathbf{t}]$, which relate the camera orientation and position to a world coordinate system, are called the extrinsic parameters. $\mathbf{K}$ and $[\mathbf{R}|\mathbf{t}]$ can be obtained by the calibration methods.

The presented camera imaging model is ideal and does not account for lens distortion. Therefore, the radial distortion and tangential distortion models are used to accurately represent a real camera:

(1) Radial distortion model:

$$\begin{aligned} x_{\text{distorted}} &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ y_{\text{distorted}} &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \end{aligned} \quad (2)$$

where k_1 , k_2 and k_3 are parameters of the radial distortion model; $r^2 = x^2 + y^2$. ($x_{\text{distorted}}$, $y_{\text{distorted}}$) and (x, y) are the normalized image coordinates.

(2) Tangential distortion model:

$$\begin{aligned} x_{\text{distorted}} &= x + 2p_1 xy + p_2(r^2 + 2x^2) \\ y_{\text{distorted}} &= y + p_1(r^2 + 2y^2) + 2p_2 xy \end{aligned} \quad (3)$$

where p_1 and p_2 are the parameters of the tangential distortion model. By combining Equations (2) and (3), it can be concluded that:

$$\begin{aligned} x_{\text{distorted}} &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_1 xy + p_2(r^2 + 2x^2) \\ y_{\text{distorted}} &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + p_1(r^2 + 2y^2) + 2p_2 xy \end{aligned} \quad (4)$$

$$\begin{aligned} u &= x f_x + u_0 \\ v &= y f_y + v_0 \end{aligned} \quad (5)$$

Thus, for any point P , the correct position of this point (u, v) on the pixel plane can be found through five distortion coefficients $(k_1, k_2, k_3, p_1, p_2)$.

4.2. Ray Intersection with the DSM

According to Equation (1), when a camera is calibrated, $\mathbf{C}$ is determined. Thus, we can map points (X_W, Y_W, Z_W) in space to points (u, v) on the image plane under the camera model. On the contrary, we cannot calculate 3D coordinates from image pixels, because

depth information is lost during the imaging process; that is, we cannot recover 3D from 2D. Therefore, it is necessary to use topographic information. The DSM is a topographic digital model, which describes the relief and the situation on the surface. It consists of the digital terrain model and represents the Earth's surface, including all objects on it.

The DSM-based localization solution model is shown in Figure 2, where O is the camera's central point, A is the object point, and the curve L represents the terrain passing through A. All points on the imaging ray OA are imaged at the same point on the image plane. Therefore, the points on the imaging ray can be searched and matched with the DSM model. The specific calculation steps are listed as follows:

- (1) Construct an imaging ray, $(X_0, Y_0, Z_0) + k(U, V, W)$, where (X_0, Y_0, Z_0) is the camera location in space, $k \geq 0$ is an arbitrary distance, and (U, V, W) is a unit vector representing the direction of the imaging ray from the camera. As shown in Figure 2a, assume that the location of point A in the image is (u_a, v_a) , and the pixel coordinates of any point B on the imaging ray OA are also (u_a, v_a) . Substitute λ (a constant) into Equation (1) to obtain $B(X_B, Y_B, Z_B)$. This gives:

$$\begin{bmatrix} U \\ V \\ W \end{bmatrix} = \begin{bmatrix} \frac{X_B - X_0}{\sqrt{(X_B - X_0)^2 + (Y_B - Y_0)^2 + (Z_B - Z_0)^2}} \\ \frac{Y_B - Y_0}{\sqrt{(X_B - X_0)^2 + (Y_B - Y_0)^2 + (Z_B - Z_0)^2}} \\ \frac{Z_B - Z_0}{\sqrt{(X_B - X_0)^2 + (Y_B - Y_0)^2 + (Z_B - Z_0)^2}} \end{bmatrix}, \quad (6)$$

- (2) Search for object points. Start from the camera point and search along the direction of the imaging ray OA. The search step is the grid spacing Δd of the DSM. The coordinates of the N th search point are:

$$\begin{bmatrix} X_N \\ Y_N \\ Z_N \end{bmatrix} = \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} + N\Delta d \begin{bmatrix} U \\ V \\ W \end{bmatrix}. \quad (7)$$

Substitute (X_N, Y_N) into the DSM to search and match. Record the elevation at (X_N, Y_N) on the DSM as $\text{Elev}(X_N, Y_N)$. The elevations of the four corner points of the grid where (X_N, Y_N) is located are Z_1, Z_2, Z_3 and Z_4 , respectively. Then,

$$\text{Elev}(X_N, Y_N) = \frac{1}{4}(Z_1 + Z_2 + Z_3 + Z_4). \quad (8)$$

when $\text{Elev}(X_N, Y_N) \geq Z_N$ appears for the first time, it indicates that object A has passed:

- ① If $\text{Elev}(X_N, Y_N) = Z_N$, (X_N, Y_N, Z_N) is the world coordinates of object A.
- ② If $\text{Elev}(X_N, Y_N) > Z_N$, it means that the object point is located between the search point N and $N - 1$. $\text{Elev}(X_N, Y_N)$ is abbreviated as E_N . By interpolation, a more precise location estimate can be obtained. The interpolation process is shown in Figure 2b. According to the triangle proportion relation, there is:

$$\frac{\overline{Z_{N-1}A}}{\overline{Z_{N-1}E_{N-1}}} = \frac{\overline{Z_{N-1}Z_N}}{\overline{Z_N E_N} + \overline{Z_{N-1}E_{N-1}}}. \quad (9)$$

The world coordinates of object A are:

$$\begin{bmatrix} X_A \\ Y_A \\ Z_A \end{bmatrix} = \begin{bmatrix} X_{N-1} \\ Y_{N-1} \\ Z_{N-1} \end{bmatrix} + \overline{Z_{N-1}A} \begin{bmatrix} U \\ V \\ W \end{bmatrix}. \quad (10)$$

Figure 2. Ray intersection with the DSM. (a) Calculation model of the intersection's location between an imaging ray and the DSM; (b) When $\text{Elev}(X_N, Y_N) > Z_N$, there is interpolation.

4.3. Pedestrian Detection and Tracking

For surveillance video in fixed scenes, people pay more attention to the moving objects without concern for the background information. Therefore, the moving objects are also the emphasis in video intelligent analysis.

You Only Look Once (YOLO) is a family of famous models for being highly performant yet incredibly small. YOLOv5 is lightweight and has good advantages in detecting small objects while considering accuracy and speed. In addition, the anchor box selection process is integrated into YOLOv5. Thus, without considering any of the datasets as input, it will automatically “learn” to provide the best anchor boxes for that dataset and use them during training. Figure 3 shows the network architecture of YOLOv5. It consists of three main architectural blocks: (i) backbone, (ii) neck, and (iii) head. The focus structure is added in the backbone, and two CSP structures are designed. The CSP1-x structure is incorporated in DarkNet, creating CSPDarknet which is the backbone of YOLOv5. CSPDarknet extracts features from images consisting of CSP1-x networks. The CSP solves duplicate gradient problems in large-scale backbones resulting in fewer parameters and fewer FLOPS (floating-point operations per second). In turn, it ensures the inference speed and accuracy and reduces the model size. The neck part uses PANet to generate a feature pyramids network to perform aggregation on the features and pass it to the Head for prediction. The CSP2-X structure used in the neck strengthens the capability of network feature fusion. The head of YOLOv5 generates three different sizes of feature maps for multi-scale prediction. The detection results include class, score, location, and size.

The method of deep SORT [40] adopts recursive Kalman filtering and frame-by-frame data association to realize object tracking. The Mahalanobis distance is used to incorporate motion information, and the smallest cosine distance is used to associate the appearance information of objects. Using a weighted sum according to Equation (11), the above two metric measures are combined to confirm the final metric and to output object tracking information through the Hungarian algorithm to achieve matching:

$$c_{i,j} = \lambda d^{(1)}(i,j) + (1 - \lambda) d^{(2)}(i,j), \quad (11)$$

where $d^{(1)}(i,j)$ is the (squared) Mahalanobis distance, $d^{(2)}(i,j)$ is the cosine distance and λ is a hyperparameter.

We adopt YOLOv5 to filter out every detection that is not a person and then use a deep SORT algorithm to track the persons who are detected by YOLOv5. That is because the deep association metric is trained on a person-only dataset. The output results include (frame ID, u_l , v_u , u_r , v_d) (Figure 4). (u_l , v_u , u_r , v_d) describe the size and position of the bounding box.

Figure 3. The network architecture of YOLOv5. (1) Backbone: CSPDarknet for feature extraction; (2) Neck: PANet for feature fusion; (3) Head: YOLO layer for prediction.

Figure 4. Results of object tracking. Numbers 1, 2, 3, 4, 5 are object IDs in this image. The size and position of the bounding box are (u_l, v_u, u_r, v_d) .

Because this paper locates objects by intersecting the imaging ray and the DSM, it is reasonable to choose the center point of the standing position on both feet. Pedestrians are in motion most of the time. Many object frames are used to measure the position of the center point of both feet. It is reasonable to calculate according to the following equation:

$$\begin{aligned} u_f &= \frac{u_l + u_r}{2} \\ v_f &= v_d + \frac{10(v_d - v_u)}{11} \end{aligned} \quad (12)$$

Obtain object tracking results according to Equation (12) and correct them according to the distortion correction model of the camera. Then, the distortion correction results are transferred to the mapping model to obtain the geographical trajectories of the objects.

4.4. Acquisition of Objects' Spatial–Temporal Information

Assuming that the pedestrian is perpendicular to the ground, and the world coordinates of the contact point of the center of the feet are (X_a, Y_a, Z_a) , (X_a, Y_a) is therefore the pedestrian's head location. The position of the head is the center point of the upper edge of the bounding box, which is denoted as (u_t, v_t) . Substituting the camera imaging model (Equation (1)), the elevation of the head is solved as Z_t . Therefore, the height of the pedestrian is:

$$H = Z_t - Z_a. \quad (13)$$

The body width of the object is calculated by the width of the bounding box. The two points where the horizontal line containing the center point of both feet meets the object bounding box are (u_l, v_f) and (u_r, v_f) , respectively. (u_l, v_f) , (u_r, v_f) and the elevation Z_a are substituted into the camera imaging model, that the corresponding world coordinates can be obtained as (X_l, Y_l, Z_a) and (X_r, Y_r, Z_a) . The width of the object is:

$$W = \sqrt{(X_l - X_r)^2 + (Y_l - Y_r)^2}, \quad (14)$$

Using the object ID as the unique identifier, the spatial–temporal information of objects, such as the 3D geographic coordinates, object height, width, and frame ID, are stored.

5. Experiments and Results

5.1. Experimental Environment

We first proposed a general fusion framework of 3D geographic information and moving objects in surveillance video. Our focus is to obtain the 3D spatial–temporal information of the moving objects and display their trajectories in the 3D model. As far as we know, there are no publicly available datasets for this research topic. We cannot compare our results with state-of-the-art. In order to verify our experiment, we chose a corner of an office park, which is divided into left and right areas bounded by steps. Although both the right and left areas are relatively flat, the two areas are not in the same plane. The elevation difference between the two areas is between 50–60 cm. The right area was selected for the experiment. There are very clear tile gaps in the right area, which are convenient for measuring the true value of the moving object trajectories and to compare the object positioning trajectories with the real trajectories. At the same time, it is more convenient for readers to see the complete visualization results and intuitively judge the positioning effects.

In the experiment, a model DH-SD-6C3230U-HN-D2 200 W network core high-speed intelligent white light dome camera (Dahua) was used. Its video resolution was 1920×1080 , and the frame rate was 25 fps. A 36-s surveillance video with a total of 900 frames was taken. The 3D model was obtained by photogrammetry. The coordinate system used was the China Geodetic Coordinate System 2000, and the elevation system was the GPS geodetic height, which provided a reference frame. The geographical coordinates of the study area were large and the charts were not concise and easy to read. At the same time, considering the confidentiality of the surveying results, the data used in this paper were translated and rotated.

5.2. Intrinsic Parameters

As shown in Figure 5, an alumina model plane with an array of 12×9 , a square side length of 40 mm, and an accuracy of 0.01 mm was used. A few images of the model plane

under different orientations were taken by moving the plane. We used these images to calibrate this camera by the Zhang Zhengyou calibration method [14]. To obtain a higher calibration accuracy, the images with large average reprojection errors were deleted. Finally, 15 images were chosen for calibration. The results of the calibration are as follows:

Figure 5. Eight images of the chessboard.

The intrinsic matrix is given by: $K = \begin{bmatrix} 1581.5766 & -0.5230 & 1012.4398 \\ 0 & 1581.1851 & 559.7023 \\ 0 & 0 & 1 \end{bmatrix}$.

The radial distortion is [0.0314 0.0358 0.2725]. The tangential distortion is [0.0004 −0.0002], and the reprojection error is 0.13 pixels.

5.3. Extrinsic Parameters

Five landmark points were evenly distributed within the camera's viewing range, and a total station was used to measure the three-dimensional coordinates of the five landmarks. Because of the distortion of the camera, it will bring some errors to the extrinsic parameters. As a result, first the distortion correction was completed, and then the pixel coordinates of the five landmark points were collected, which are shown in Figure 6. Via the five pairs of points, the EPNP method [18] was proved to calculate the extrinsic parameter matrix of the camera. Corresponding point pairs are shown in Table 1.

(a)

(b)

Figure 6. Measuring the three-dimensional coordinates of landmarks. (a) Before the distortion correction; (b) after the distortion correction.

Table 1. Point pairs corresponding to image coordinates and world coordinates.

Point	u/pixel	v/pixel	X/m	Y/m	Z/m
B1	1220	848	46.129	75.967	87.433
B6	472	713	42.334	69.258	87.127
B25	700	401	50.669	53.49	87.186
B32	756	242	68.291	14.321	87.586
B37	1193	219	56.517	62.117	90.979

The coordinates of the calibrated camera in the world coordinate system are (39.314, 84.267, 92.512). The rotation matrix and translation matrix are as follows:

$$R = \begin{bmatrix} -0.851808752945461 & -0.523751350419021 & 0.010313648226996 \\ 0.146042440299173 & -0.218518158930861 & 0.964842691763237 \\ -0.503083943330495 & 0.823367680414843 & 0.262625605780016 \end{bmatrix}$$

$$t = \begin{bmatrix} 76.6676929041603 \\ -76.5872215961609 \\ -73.8995528853148 \end{bmatrix}$$

So far, $\mathbf{K}$, $\mathbf{R}$, $\mathbf{t}$, and $(k_1, k_2, k_3, p_1, p_2)$ have been obtained, and the imaging model of the camera has been determined.

5.4. Object Tracking

The proposed method was adopted to detect and track pedestrians in the surveillance video, and some original frames and tracking frames were selected (Figure 7). As can be seen from Figure 7, there is no matching error for the five moving objects, and the tracking results are good.

Figure 7. The corresponding frame of the original video and the tracking video. From left to right, the five images in the first row are the 200th, 400th, 600th and 800th frames of the original video. The five images in row 2 are the 200th, 400th, 600th, and 800th frames of the tracking video.

5.5. Estimating the Object Location

Each location is obtained by shooting a 3D imaging ray through the center of the camera and the center point of both feet on the image plane into the scene and determining where it intersects with the terrain. The DSM represents the most realistic representation of the topography, which is obtained through the 3D model. The DSM grid is $5 \text{ cm} \times 5 \text{ cm}$ with an accuracy of 6.5 cm (only the ground). Because the spacing of the DSM grid is less than its elevation accuracy, object locations are determined by using $\text{Elev}(X_N, Y_N) \geq Z_N$ as the decision condition, without interpolation processing in this experiment.

Figure 8 shows the planar trajectories of moving objects in the world coordinate system. Although there are some jitters, they reflect the actual motion of the moving objects as a whole. Figure 9 gives the elevation trajectories. In Figure 9a the elevations of the first, third, fourth, and fifth objects are gentle, and there is no sharp fluctuation. However, the elevation of the second object jitters. The second object moves close to the flower bed. Affected by the errors, the center points of the feet in some frames are mapped to points on the flower bed, not on the ground, which is not reasonable. To accurately reflect the elevation information of moving objects, median filtering is adopted. The elevations are filtered by median filtering with a window size of 25. The spikes in Figure 9b can be filtered out by median filtering. Then, the trajectories are fitted by a cubic polynomial in Figure 9c. Satisfactory results have been achieved in applying the cubic polynomial to the fitting of elevation trajectories based on the computation principle of the least square method. The three-dimensional trajectories of moving objects are presented in a graph (Figure 10) in which the elevations are filtered by median filtering first and then fitted by a cubic polynomial.

Figure 8. The planar trajectories of moving objects in the world coordinate system in which the ordinate is X (m), and the abscissa is Y (m).

Figure 9. The elevation trajectories of moving objects in the world coordinate system. (a) The original trajectories in elevation; (b) the trajectories in elevation filtered by median filtering with the window size of 25; (c) the trajectories in elevation filtered by a cubic polynomial. The ordinate is Z (m), and the abscissa is the video frame number in (a–c).

Figure 10. The 3D trajectories of moving objects in the world coordinate system in a graph. The planar positions are the mapping results, and the elevations are the results first filtered by median filtering and then fitted by a cubic polynomial.

5.6. Estimating the Object Width and Height

In Section 5.5, we have known object locations in each frame of the sequence. Meanwhile, we also want to know the object width and height. Equations (13) and (14) give a method to calculate the object width and height. Due to the influence of walking postures, the calculation results of the object width and height will not be fixed. Thus, the median filtering with a window size of 25 is also used here to filter object width and height.

Figures 11 and 12 show the width and height of moving objects, respectively. After removing anomalies with the median filtering, the average width and height are calculated in Equation (15):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (15)$$

where x_i is the object width/height after filtering in each frame and n is the number of frames. The average width and height of moving objects are shown in Table 2.

Figure 11. The width of moving objects. (a) The original calculation results of the object width; (b) the calculation results of the object width filtered by median filtering. The ordinate is the width (m), and the abscissa is the video frame number.

Figure 12. The height of moving objects. (a) The original calculation results of the object height; (b) the calculation results of the object height filtered by median filtering. The ordinate is the height (m), and the abscissa is the video frame number.

Table 2. Average width and height of moving objects.

Object ID	1	2	3	4	5
Average width/m	0.59	0.64	0.66	0.58	0.62
Average height/m	1.59	1.76	1.75	1.61	1.80

5.7. Spatial–Temporal Information of Moving Objects

We have obtained the spatial–temporal information of moving objects, including 3D world coordinates, width, and height. Table 3 gives part of the spatial–temporal information of the fourth object, which shows the location, width, and height of the fourth object in the geographic scene from the 368th frame to the 384th frame.

Table 3. Part of the spatial–temporal information of the fourth object.

Object ID	X/m	Y/m	Z/m	Width/m	Height/m	Frame
4	53.323	54.572	87.267	0.603	1.622	368
4	53.367	54.480	87.266	0.603	1.612	369
4	53.367	54.480	87.266	0.603	1.612	370
4	53.386	54.489	87.266	0.603	1.608	371
4	53.386	54.489	87.266	0.603	1.608	372
4	53.341	54.581	87.267	0.603	1.608	373
4	53.360	54.590	87.267	0.594	1.608	374
4	53.379	54.599	87.267	0.594	1.608	375
4	53.156	55.109	87.270	0.594	1.608	376
4	53.156	55.109	87.270	0.594	1.608	377
4	53.156	55.109	87.270	0.594	1.608	378
4	53.192	55.126	87.270	0.594	1.608	379
4	53.277	54.999	87.262	0.592	1.608	380
4	53.274	55.052	87.265	0.585	1.608	381
4	53.292	55.061	87.265	0.585	1.626	382
4	53.358	54.925	87.264	0.585	1.628	383
4	53.377	54.934	87.264	0.585	1.631	384

5.8. Statistics of the Experimental Time

We conducted experiments in CPU and GPU hardware environments, respectively. The duration of the surveillance video used in the experiments is 36 s, with 25 frames

per second and 900 frames in total. The experimental results are shown in Table 4. In the current experimental design, we did not perform operations such as keyframe extraction on the video, and this part will be studied in the follow-up.

Table 4. The experimental time statistics in CPU and GPU hardware environments, respectively.

Core Hardware for Computing	CPU	GPU
Model	Intel Xeon W-2145 3.70 GHz 8 G RAM	Nvidia Quadro P4000
Object Detection and Tracking Time (s)	1068.37	125.84
Spatial-temporal Information Extraction Time (s)	0.84	0.84
Total Computing Time (s)	1069.21	126.68
Average Computing Time per Frame (s)	1.19	0.14

5.9. Analysis of the Experimental Results

The walking routes of pedestrians were measured by a total station to verify the accuracy of the experimental results, which was used as the true value of the trajectories. The comparison and analysis of the measured and mapped trajectories are shown in Figure 13.

Figure 13. Comparison of the planar position between the measured and mapped trajectories in which the ordinate is X (m), and the abscissa is Y(m).

The vertical distance between the mapping point and the measured trajectory is taken as the standard for evaluating the error of the planar position. The maximum error (ME) and root mean square error (RMSE) of the dynamic object position are:

$$ME = \text{Max}(d_1, d_2 \dots, d_n), \quad (16)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}}, \quad (17)$$

where d_i is the vertical distance between the i th mapping point and the measured trajectory, and n is the number of the mapping points. The ME and RMSE of the planar position and elevation are shown in Tables 5 and 6, respectively.

Table 5. Planar errors of the moving objects.

Object ID	1	2	3	4	5
ME/cm	31	40	22	48	19
RMSE/cm	4	6	5	7	4

Table 6. Elevation errors of the moving objects.

Object ID	1	2	3	4	5
ME/cm	9	10	10	9	9
RMSE/cm	1.9	2.2	2	1.7	2.3

As can be seen from Figure 13, the planar positions of the five objects all vibrate. It is because a person is not a rigid body. In the movement process, the bounding box will fluctuate due to the influence of arm swinging, leg stepping and turning, etc. In particular, the fourth object turns around frequently near the starting position, causing severe shaking near the starting section. Meanwhile, the second and fourth objects also show large fluctuations on the far right of Figure 13. The reason is that the second and fourth objects are about to be away from the field of view of the camera and blocked by the nearby flower bed. The bounding box is not precisely accurate. The ME of the second and fourth objects are 40 cm and 48 cm (Table 5), respectively, which is larger than those of the first, third, and fifth objects. In other positions, the ME of the second and fourth objects are 25 cm and 28 cm, respectively. The RMSE of the five objects is relatively small, at less than 10 cm. As a whole, in the case that the bounding box can thoroughly outline the entire object, the approach proposed in this paper can realize the positioning of moving objects with high accuracy, and the ME of the planar position can be controlled within 31 cm and the RMSE within 10 cm.

As shown in Figure 14, with the increase in the number of frames, all of the elevation errors show a decreasing trend, which is mainly affected by the accuracy of the DSM. The monitoring area is located between two tall buildings, and the objects are moving from the middle of the two tall buildings to the open area outside. Affected by satellite signals, the accuracy of the data acquisition of the 3D model is poor in the middle area and better in the outer open area, resulting in the uneven accuracy of the DSM. The data in Figure 14 are statistically analyzed to obtain the ME and RMSE of each object (Table 6). It can be seen from Table 6 that the ME of the five objects in elevation are all within 10 cm, and the RMSE are all within 3 cm, which has a high elevation accuracy.

Figure 14. Comparison in elevation between the measured and mapped trajectories in which the ordinate is Z (m), and the abscissa is the video frame number.

We have evaluated the motion trajectories of moving objects. The method proposed in this paper has achieved good results in both planar accuracy and elevation accuracy. Next, we evaluate the geometric information of moving objects. The first step is to measure the widest position of a body with feet shoulder-width apart and arms hanging down naturally. Then, the measured results are taken as the true value of the object width. The true value of the object height is acquired in the usual way. Stand straight with heels together and measure the height from the heels to the top of the head. We compared the calculated width and height with the true value in Table 7, where w and h are the calculated results from Table 2, and $\tilde{w}$ and $\tilde{h}$ are the true values of width and height, respectively. The error is equal to the calculated value minus the true value. From the comparison results, the errors

in height are all small (within 2 cm). The errors in width are slightly larger than the errors in height (within 5 cm).

Table 7. Errors in the geometry information of the moving objects.

Object ID	Width/m		Error/cm	Height/m		Error/cm
	w	$\tilde{w}$		h	$\tilde{h}$	
1	0.59	0.55	4	1.59	1.61	−2
2	0.64	0.59	5	1.76	1.75	1
3	0.66	0.61	5	1.75	1.73	2
4	0.58	0.54	4	1.61	1.60	1
5	0.62	0.60	2	1.80	1.78	2

6. Visualization

We mapped moving objects in surveillance video to three-dimensional geographic space, which can not only overcome the disadvantage of video redundancy, but also be beneficial for managers to monitor objects more intuitively, and facilitate the realization of measurement, statistics, and the analysis of objects. The 3D geographic information model used in this paper was obtained by Dajiang unmanned aerial vehicle photogrammetry. The root mean square error of the planar position and evaluation is used as the standard of accuracy. The corresponding equations are:

$$\text{RMSE}_{(x,y)} = \sqrt{\frac{\sum_{i=1}^n (x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2}{n}}, \quad (18)$$

$$\text{RMSE}_z = \sqrt{\frac{\sum_{i=1}^n (z_i - \tilde{z}_i)^2}{n}}, \quad (19)$$

where (x_i, y_i, z_i) are the coordinates in the 3D model, $(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)$ are the coordinates measured in the field, and n is the number of points. Equation (18) is the calculation equation of the planar error, and Equation (19) is the calculation equation of the evaluation error. Without considering flower beds, trees, houses, and other situations, 66 points on the ground were measured by a total station in the field, and the corresponding point coordinates were taken from the 3D model in the office. After calculation, the planar accuracy of the 3D model was 3.7 cm, and the evaluation accuracy was 6.5 cm. The 3D model has high accuracy.

The motion trajectories of the moving objects in the surveillance video emerged in the 3D model, as shown in Figure 15. We provided the trajectories of the objects in the 3D model from the side view, top view, and close to the camera's direction.

In Figure 15, we can intuitively see that all of the five objects walk along the gaps among the floor tiles. The first and fourth objects make two right-angle turns; the second, third and fifth objects all walk straight. As shown in Figure 15b, c, the five objects' trajectories are identical to the actual walking routes. In Figure 15a, the object trajectories can be found to perch on the ground, and there is no suspension, which indicates that the elevation error is small. Figure 15 shows the consistency between the mapped trajectories and the actual trajectories, which fully demonstrates that the method in this paper can extract objects' spatial-temporal information and has a high positioning accuracy.

Figure 15. Display of multi-object trajectories in the 3D model from three different views. (a) Side view; (b) top view; (c) roughly the camera's direction.

7. Conclusions

The goal of this paper was to realize the fusion of 3D geographic information and moving objects. The fusion paves the way for new opportunities, not only for computer vision but also for geomatics. It can help users to understand a video in a unified geographical framework. The ability to acquire the geolocation of each moving object in the video relies on the quality of the mapping model, which is closely related to the camera model, object pixel coordinates, and the DSM accuracy. In this paper, the camera model was determined by camera calibration, the extraction of moving objects was obtained by YOLOv5 and deep SORT, and the DSM was acquired by a 3D geographic model. After distortion correction, pixel coordinates of moving objects were sent to the mapping model. We knew the moving objects with the geospatial information and then calculated the object width and height. At last, the trajectories of moving objects were presented in a 3D geographical scene. To verify the effectiveness of the proposed method, the experimental results were compared with the true values. The comparison results show that the proposed method has achieved very good accuracy in geographic location and the geometric measurement of moving objects in surveillance video.

The framework proposed in this paper is of great significance and provides ideas for related research. We parsed the pixel coordinates into 3D geographic coordinates to achieve the accurate positioning and measurement of monitoring objects, which provides favorable technical support for urban security, including objects' spatial-temporal analysis, object search, abnormal alarms, and object statistics. At the same time, it is also a sub-subject of 3DCM application. Some scholars have begun to study life expectancy by measuring people's walking speed. There is no doubt that the spatial-temporal data we obtained can provide analytical data for these scholars. Based on the importance of this research topic,

we will study the integration of multi-camera and multi-object expression in a unified 3D geographical scene.

Author Contributions: Guarantor: Shijing Han, Xiaorui Dong, Xiangyang Hao and Shufeng Miao are responsible for the entirety of the work and the final decision to submit the manuscript; Conceptualization and methodology, Shijing Han and Xiangyang Hao; data acquisition, processing, and analysis, Shijing Han, Shufeng Miao and Xiaorui Dong; software, Shufeng Miao and Xiaorui Dong; writing—original draft preparation, Shijing Han; writing—review and editing, Shijing Han and Xiangyang Hao; funding acquisition, Xiangyang Hao. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the university research team development fund, grant number f4211.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to shijinghan@nnn.edu.cn.

Acknowledgments: Thanks to Long Yang for his technical support in producing the 3D model and DSM. Thanks to Jinsheng Liu, Anning Liu, and Xiaoping Song for their cooperation in acquiring surveillance videos and field surveying.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Vis. Commun. Image Represent.* **2021**, *77*, 103116. [\[CrossRef\]](#)
2. Lee, S.C.; Nevatia, R. Robust camera calibration tool for video surveillance camera in urban environment. In Proceedings of the CVPR 2011 Workshops, Colorado Springs, CO, USA, 20–25 June 2011; pp. 62–67.
3. Eldrandaly, K.A.; Abdel-Basset, M.; Abdel-Fatah, L. PTZ-surveillance coverage based on artificial intelligence for smart cities. *Int. J. Inf. Manag.* **2019**, *49*, 520–532. [\[CrossRef\]](#)
4. Liu, S.; Liu, D.; Srivastava, G.; Połap, D.; Woźniak, M. Overview and methods of correlation filter algorithms in object tracking. *Complex Intell. Syst.* **2020**, *7*, 1895–1917. [\[CrossRef\]](#)
5. Kawasaki, N.; Takai, Y. Video monitoring system for security surveillance based on augmented reality. In Proceedings of the 12th International Conference on Artificial Reality and Telexistence, Tokyo, Japan, 4–6 December 2002; pp. 4–6.
6. Sankaranarayanan, K.; Davis, J.W. A fast linear registration framework for multi-camera GIS coordination. In Proceedings of the 2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, Santa Fe, NM, USA, 1–3 September 2008; pp. 245–251.
7. Xie, Y.; Wang, M.; Liu, X.; Mao, B.; Wang, F. Integration of Multi-Camera Video Moving Objects and GIS. *ISPRS Int. J. Geo. Inf.* **2019**, *8*, 561. [\[CrossRef\]](#)
8. Xie, Y.; Wang, M.; Liu, X.; Wu, Y. Integration of GIS and moving objects in surveillance video. *ISPRS Int. J. Geo. Inf.* **2017**, *6*, 94. [\[CrossRef\]](#)
9. Girgensohn, A.; Kimber, D.; Vaughan, J.; Yang, T.; Shipman, F.; Turner, T.; Rieffel, E.; Wilcox, L.; Chen, F.; Dunnigan, T. Dots: Support for effective video surveillance. In Proceedings of the 15th ACM international conference on Multimedia, Bavaria, Germany, 24–29 September 2007; pp. 423–432.
10. Han, L.; Huang, B.; Chen, L. Integration and application of video surveillance system and 3D GIS. In Proceedings of the 2010 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010; pp. 1–5.
11. Zheng, J.; Zhang, D.; Zhang, Z.; Lu, X. An integrated system of video surveillance and GIS. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *170*, 022088. [\[CrossRef\]](#)
12. Xie, Y.; Wang, M.; Liu, X.; Wu, Y. Surveillance Video Synopsis in GIS. *ISPRS Int. J. Geo Inf.* **2017**, *6*, 333. [\[CrossRef\]](#)
13. Song, H.; Liu, X.; Zhang, X.; Hu, J. Real-time monitoring for crowd counting using video surveillance and GIS. In Proceedings of the 2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering, Nanjing, China, 1–3 June 2012; pp. 1–4.
14. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [\[CrossRef\]](#)
15. Hartley, R.I. Self-calibration of stationary cameras. *Int. J. Comput. Vis.* **1997**, *22*, 5–23. [\[CrossRef\]](#)
16. Triggs, B. Autocalibration and the absolute quadric. In Proceedings of the Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; pp. 609–614.

17. Lu, X.X. A review of solutions for perspective-n-point problem in camera pose estimation. *J. Phys. Conf. Ser.* **2018**, *1087*, 052009. [\[CrossRef\]](#)
18. Lepetit, V.; Moreno-Noguer, F.; Fua, P. Epnp: An accurate o (n) solution to the pnp problem. *Int. J. Comput. Vis.* **2009**, *81*, 155. [\[CrossRef\]](#)
19. Li, S.; Xu, C.; Xie, M. A robust O (n) solution to the perspective-n-point problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1444–1450. [\[CrossRef\]](#)
20. Collins, R.; Tsin, Y.; Miller, J.R.; Lipton, A. Using a DEM to determine geospatial object trajectories. In Proceedings of the DARPA Image Understanding Workshop, Monterey, CA, USA, 20–23 November 1998; pp. 115–122.
21. Milosavljević, A.; Rančić, D.; Dimitrijević, A.; Predić, B.; Mihajlović, V. Integration of GIS and video surveillance. *Int. J. Geogr. Inf. Sci.* **2016**, 1–19. [\[CrossRef\]](#)
22. Milosavljević, A.; Rančić, D.; Dimitrijević, A.; Predić, B.; Mihajlović, V. A Method for Estimating Surveillance Video Georeferences. *ISPRS Int. J. Geo. Inf.* **2017**, *6*, 211. [\[CrossRef\]](#)
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
24. Girshick, R. Fast r-cnn. In Proceedings of the Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
25. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 91–99. [\[CrossRef\]](#) [\[PubMed\]](#)
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the Proceedings of the IEEE international Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
27. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
28. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
29. Huang, C.; Li, Y.; Nevatia, R. Multiple target tracking by learning-based hierarchical association of detection responses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 898–910. [\[CrossRef\]](#) [\[PubMed\]](#)
30. He, Y.; Wei, X.; Hong, X.; Shi, W.; Gong, Y. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Trans. Image Processing* **2020**, *29*, 5191–5205. [\[CrossRef\]](#)
31. Xu, J.; Bo, C.; Wang, D. A novel multi-target multi-camera tracking approach based on feature grouping. *Comput. Electr. Eng.* **2021**, *92*, 107153. [\[CrossRef\]](#)
32. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 107–122.
33. Ristani, E.; Tomasi, C. Features for multi-target multi-camera tracking and re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6036–6046.
34. Zhang, Z.; Wu, J.; Zhang, X.; Zhang, C. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *arXiv* **2017**, arXiv:1712.09531.
35. Tagore, N.K.; Singh, A.; Manche, S.; Chattopadhyay, P. Deep Learning based Person Re-identification. *arXiv* **2020**, arXiv:2005.03293.
36. Katkore, A.; Moezzi, S.; Kuramura, D.Y.; Kelly, P.; Jain, R. Towards video-based immersive environments. *Multimed. Syst.* **1997**, *5*, 69–85. [\[CrossRef\]](#)
37. Takehara, T.; Nakashima, Y.; Nitta, N.; Babaguchi, N. Digital diorama: Sensing-based real-world visualization. In Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Dortmund, Germany, 28 June–2 July 2010; Springer: Berlin/Heidelberg, Germany, 2010; Volume II, pp. 663–672.
38. Zhang, X.; Liu, X.; Song, H. Video surveillance GIS: A novel application. In Proceedings of the 2013 21st International Conference on Geoinformatics, Kaifeng, China, 20–22 June 2013; pp. 1–4.
39. Yang, Y.; Chang, M.-C.; Tu, P.; Lyu, S. Seeing as it happens: Real time 3D video event visualization. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec, ON, Canada, 27–30 September 2015; pp. 2875–2879.
40. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.