International Journal of Geo-Information

MDPI

Article

An Accurate Matching Method for Projecting Vector Data into Surveillance Video to Monitor and Protect Cultivated Land

Zhenfeng Shao [1], Congmin Li [1,*], Deren Li [1], Orhan Altan [2], Lei Zhang [3] and Lin Ding [3]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; shaozhenfeng@whu.edu.cn (Z.S.); Drli@whu.edu.cn (D.L.)
[2] Department of Geomatics Engineering, Istanbul Technical University, Istanbul 36626, Turkey; oaltan@itu.edu.tr
[3] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; zhanglei1990@whu.edu.cn (L.Z.); dinglin@whu.edu.cn (L.D.)
* Correspondence: cminlee@whu.edu.cn

Received: 25 May 2020; Accepted: 15 July 2020; Published: 17 July 2020

Abstract: The integration of intelligent video surveillance and GIS (geograhical information system) data provides a new opportunity for monitoring and protecting cultivated land. For a GIS-based video monitoring system, the prerequisite is to align the GIS data with video image. However, existing methods or systems have their own shortcomings when implemented in monitoring cultivated land. To address this problem, this paper aims to propose an accurate matching method for projecting vector data into surveillance video, considering the topographic characteristics of cultivated land in plain area. Once an adequate number of control points are identified from 2D (two-dimensional) GIS data and the selected reference video image, the alignment of 2D GIS data and PTZ (pan-tilt-zoom) video frames can be realized by automatic feature matching method. Based on the alignment results, we can easily identify the occurrence of farmland destruction by visually inspecting the image content covering the 2D vector area. Furthermore, a prototype of intelligent surveillance video system for cultivated land is constructed and several experiments are conducted to validate the proposed approach. Experimental results show that the proposed alignment methods can achieve a high accuracy and satisfy the requirements of cultivated land monitoring.

Keywords: digital orthophoto map (DOM); cultivated land protection; video GIS; basic farmland protection zones (BFPZs)

1. Introduction

China is experiencing an unprecedented urbanization process, with the urbanization level increasing from 17.92% in 1978 to 59.58% in 2018 [1]. With the acceleration of urbanization, the expansion of urban and population growth has led to an increasing demand for cultivated land, resulting in unreasonable land use. Large amounts of farmland have been occupied and converted into other uses, such as residential, industrial, commercial, infrastructure, and institutional uses during the period of urban expansion [2]. Meanwhile, idleness of cultivated land is pervasive in rural areas as a consequence of urbanization or unclear ownership. As China is a developing country with a large population and scarce land resources, the loss of cultivated land will threaten the security of food supply and the stability of society. In this respect, accurate monitoring of cultivated land plays a crucial role in farmland preservation.

In order to ensure the basic food supply for human survival, Chinese government has formulated a series of policies to protect the total amount of cultivated land, such as the delineation of basic

farmland protection zones (BFPZs) [3–5]. When a piece of land is determined as basic farmland by the land management department, it means that the land use type within the delimited boundaries cannot be changed into non-cultivated land use. If changed, it will be regarded as illegal occupation of farms. Therefore, the focus of cultivated land monitoring is to check whether there is deduction of farms and change in land use type in the delimited area over a period.

Generally, the methods used for monitoring cultivated land can be classified into two categories: field survey and remote sensing technologies. Traditional field survey method can provide more accurate results than other approaches. However, it requires considerable manpower, which can be time consuming and expensive. For the methods of using satellite remote sensing, emerging studies are mainly focused on landform classification, which is an effective way for dynamically monitoring land use and land cover changes. It is noted that Landsat series of images are widely used to detect and monitor land use change, as they are freely available in forms of different spectral, resolution, and time, supporting continuous and long-term monitoring land use change in large areas [6,7]. However, the low spatial and temporal resolution of satellite images will limit its ability to detect and monitor land use in small sized and scattered distribution land [8]. Additionally, it is difficult to use satellite images to identify the problems of land degradation, illegal occupation, and land idleness within a certain period and time. UAVs (unmanned aerial vehicles) images seem to be an alternative remote sensing tool for the detection of land use change, due to its advantage of strong mobility and high efficiency [8–11]. Nevertheless, UAVs cannot work continuously for the monitoring of cultivated land for some reasons. UAVs are easily affected by weather conditions, complex terrain, airspace regulations, and other factors in some places. Moreover, images acquired by UAVs may contain severe geometric and radiometric distortions, which challenges the automatic processing of UAVs images and hampers the application of UAVs [12,13].

Nowadays, numerous intelligent surveillance cameras have been mounted in public places due to its increasingly significant role in traffic monitoring, public security, and other applications. Intelligent video surveillance can automatically detect, track, and recognize interested objects, and analyze their activities to extract useful information from collected videos [14]. Moreover, it can remotely transmit real-time images, audio, and other data to the central control room through the Internet, which offers great potential for monitoring cultivated land. In comparison with satellite remote sensing and UAVs technology, it can work continuously with higher spatial and temporal resolution with less manual intervention. What is more, it is more suitable for monitoring small size and scattered distribution land. However, there are various problems when the conventional intelligent video surveillance is directly applied to cultivated land monitoring. For example, the unauthorized conversion of certain farmland can be detected early by intelligent video surveillance, while regulars cannot quickly identify the location, including the coordinates or other semantic information of the land. When the number of surveillance cameras grows larger, it is more difficult to manage the cameras and fragmented video data [15]. More importantly, it is hard to define the boundary of cultivated land (region of interest) in every video image when the camera is panned, tilted or zoomed.

BFPZs delimited by the land management department define the boundary of protected cultivated land, and it is usually presented and stored in forms of vector data. The integration of intelligent video surveillance and 2D vector BFPZs data provides a new opportunity for solving the above problems. As video is composed of a sequence of separate frames (images), it is essential to establish the relations that map the frame's pixels with corresponding geographic locations for each frame to integrate GIS and video [16]. At present, the general method is to directly project video frames into a virtual 3D (three-dimensional) GIS scene, which requires both precise 3D models of the surroundings and camera pose estimation [15–17]. In this case, video frames do not match the intuitive feeling of the human eye because they have been geometrically corrected to the 3D GIS scene. Furthermore, it is difficult to obtain accurate real-world physical information of acquired images due to changes in focal length and angles of the PTZ (pan-tilt-zoom) cameras. Moreover, high resolution DEM (digital elevation model) or DSM (digital surface model) are not always accessible. Another method is based on a

homography transformation, which assumes a planar ground in geographic space and requires at least four corresponding points to calculate the homography matrix parameters [18]. Both of the two methods rely on matching corresponding points in video image space and geospatial space [19]. It is a challenge to automatically match features from video image and GIS data. For PTZ cameras, it is not advisable to manually select points considering the cost and work efficiency.

Based on the above analysis, the main objective of this paper is to propose an accurate matching method for projecting vector data into surveillance video to monitor and protect the cultivated land. It mainly relies on matching the 2D image coordinates from PTZ video frames with orthophoto maps and vector surveying and mapping data. On the basis of the proposed method, we design and implement a prototype of cultivated land monitoring system. Then, the implemented prototype is applied to monitor cultivated land in Dongyang City, Zhejiang Province, China.

2. Related Work

Currently, several GIS-based video monitoring systems have been developed and applied in the fields of city safety management and forest insect defoliation and discoloration [17,18]. Inspired by these systems, this paper develops a new video and GIS integrated surveillance system for cultivated land, which can make up for the shortcomings of traditional field surveying and remote sensing technologies. In this section, we mainly introduce the related work on the integration of video and GIS.

The study mainly focuses on establishing the geometric mapping relationship between the spatial point set sampled from video frame and the geospatial point set sampled from geodetically calibrated reference imagery. Since the video is composed of a series of independent frames (images), the geometric mapping relationship should be constructed for each frame.

At present, the methods for video geo-referencing can be classified into two categories: methods based on view line intersection with DEM and methods based on a homography transformation [18–20]. For methods based on the intersection between sight and DEM, precise parameters including inner and outer camera parameters and DEM are required to project video frames into a 3D model of an area that is being watched. For example, Zhao et al. [17] proposed a video-based monitoring framework for forest insect defoliation and discoloration. GIS data and Video Surveillance could be integrated to determine the geographical position of forest insect damage in their monitoring system, which required a digital elevation model (DEM) and returned parameters from the PTZ camera. Milosavljevic et al. [19] proposed a method to estimate the geo-reference of both fixed and PTZ cameras, which relied on matching 2D image coordinates from video frames with 3D geodetic coordinates to estimate the camera's position and orientation in geographic space. In this situation, video frames do not match the intuitive feeling of the human eye because they have been geometrically corrected to the 2D or 3D GIS scene. Furthermore, it is difficult to obtain precise real-world physical information of the acquired images due to changes in focal length and angles of the PTZ cameras. Meanwhile, high resolution DEM or DSM are not always available and not all of the objects in real world can be modelled, which causes the failure of cross-mapping.

Methods based on a homography matrix assume a planar ground in a geographic space, so they require at least four matching points to estimate the corresponding homography matrix [21]. Automatically detecting and matching features from ground imagery or stationary videos, such as surveillance cameras videos, and satellite or aerial images can be defined as cross-view image matching [22–24]. It is almost impossible to automatically match features from such images due to the dramatic differences in each set of images, as they are captured from different viewpoints and in different resolutions. Therefore, the matching points are manually selected from cross-view images. The number, precision, and spatial distribution of the selected points are very important and directly affect the accuracy of image registration. For 2D GIS, the mapping from 2D geographic space to video image space is double, but for 3D GIS, the transformation is single [18]. When using these methods, the terrain model is usually taken into account for the assumption of planar ground. The main drawback of these methods is that they are unsuitable for the case of PTZ cameras. With

the change of pan-tilt-zoom, more user interaction is required to select corresponding features and prevents automation.

3. Methodology

The proposed method for integrating a surveillance video and BFPZs (2D GIS vector data) relies on registering BFPZs to surveillance video image. This can be realized by the mapping matrix estimated by matching the 2D geodetic coordinates with the 2D image coordinates in video frames. Compared to fixed camera, PTZ cameras can be controlled to rotate horizontally and vertically, and the field-of-view can be changed. The image coordinates of each video frame are independent of other images. Therefore, we need to construct the mapping relationship between BFPZs data and every video image.

As there is large resolution and geometric difference in the two types of data, it is impossible to automatically detect and match features. Therefore, we break down the integration process into two stages, as shown in Figure 1. In the configuration phase, we will determine the mapping relationship between BFPZs and the preset video image I_0. by identifying corresponding features from the two data and estimation algorithms. DOM (digital orthophoto map) is used to help identify corresponding points in the BFPZs and preset video image, as vector data is graphics and hard to be recognized. Based on these features, the transformation matrix can be estimated by algorithms. Then, the coordinates of 2D BFPZs can be transformed from geodetic space to video image space. In the dynamic mapping phase, the mapping between 2D BFPZs and PTZ video images can be seen as the problem of multi-view image matching since BFPZs has been projected in image space. Different from configuration phase, we can use a matching method to automatically detect and match points without manual intervention. Once the relative geometric relation between video frames is estimated, then the coordinates of 2D BFPZs can be transformed between different video image space. In the following sections, we will introduce the details about homography transformation, SIFT (scale-invariant feature transform) [25] and ASIFT (affine-SIFT) [26] matching algorithms and the integration strategies.

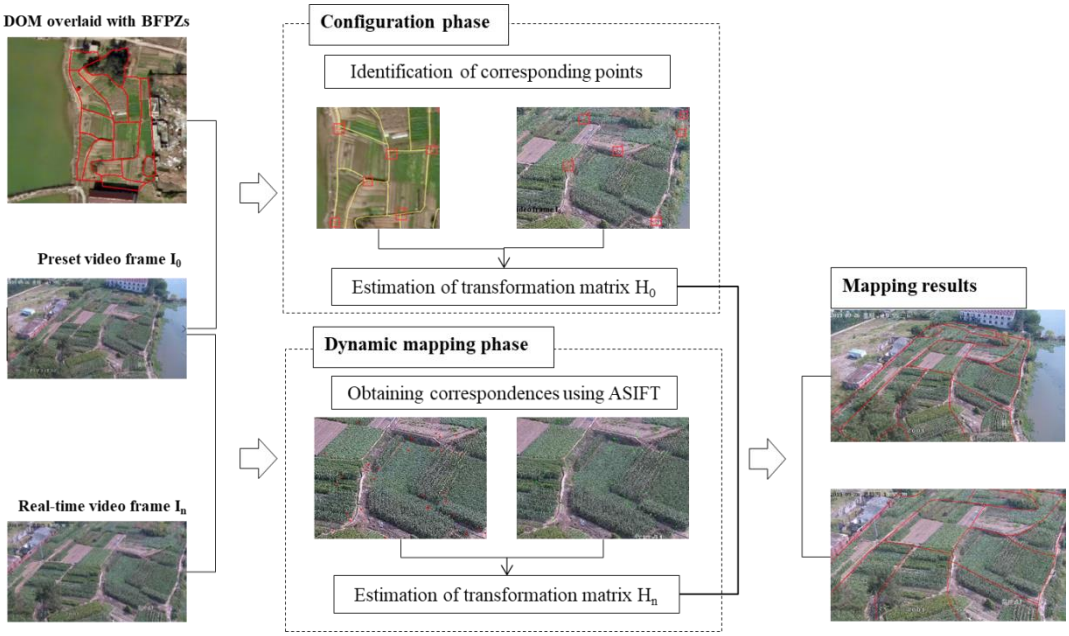


Figure 1. The flowchart of the proposed integration of 2D GIS data and PTZ (pan-tilt-zoom) surveillance video.

3.1. Homography Transformation

The homography transformation is a popular geo-referencing technique used worldwide. Surveillance video and remote sensing (RS) sensors capture the same scene differently in two views, i.e.,

front view and side view, as shown in Figure 2. Usually, the remote sensing images are geometrically corrected and used to produce a DOM. When the terrain is a planar ground or the topographic relief is small, the complex geometric relations between surveillance video image and remote sensing image can be modelled by homography transformation.

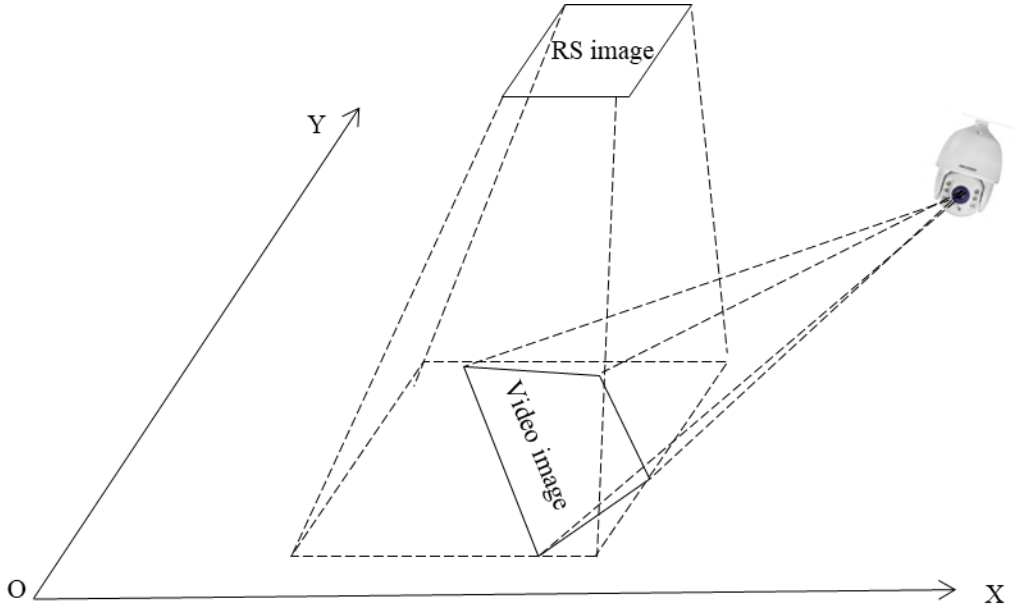


Figure 2. Different views of surveillance video and remote sensing sensors when capturing the same scene.

Assuming P is a point in the DOM spatial coordinate system, and its corresponding point in video image coordinate system is Q:

$$P = \begin{bmatrix} X & Y & 1 \end{bmatrix}^{\mathrm{T}}, \tag{1}$$

$$Q = \begin{bmatrix} u & v & 1 \end{bmatrix}^{\mathrm{T}}, \tag{2}$$

where X and Y are real world coordinates, u and v are column and row, respectively, in image coordinates.

Given a homography matrix $\mathbf{H}$, the relationship between P and Q can be expressed as follows:

$$Q = \mathbf{H}P, \tag{3}$$

$\mathbf{H}$ is a 3×3 matrix and can be represented as follows:

$$\mathbf{H} = \begin{pmatrix} A & B & C \\ D & E & F \\ G & H & 1 \end{pmatrix}, \tag{4}$$

As $\mathbf{H}$ has eight unknowns, at least four pairs of non-collinear video image point and geospatial points are required to calculate the parameters of $\mathbf{H}$. Once $\mathbf{H}$ is determined, the coordinates of any point in spatial coordinate system can be projected into the image coordinate.

$$\begin{bmatrix} u & v & 1 \end{bmatrix}^{\mathrm{T}} = \mathbf{H}\begin{bmatrix} X & Y & 1 \end{bmatrix}^{\mathrm{T}}, \tag{5}$$

The mapping is double direction, and image coordinates can also be projected into the spatial coordinate system. This modeling method is convenient for geometric presentation, computation, and implementation simplicity.

3.2. SIFT and ASIFT Matching

When the PTZ surveillance camera is paned, tilted, or zoomed, there will be geometric distortion between these video sequences. To automatically integrate GIS data and surveillance video, we use the image matching algorithm to automatically detect and match corresponding features, rather than manually selecting feature points.

SIFT [25] is a well-known image matching algorithm, which is invariant to scaling, rotation, illumination, and affine transformation with sub-pixel accuracy. The original SIFT algorithm first uses subpixel location, scale, and dominant orientation to describe a detected feature point. It can be expressed by $f = (L, \sigma, \theta, d)$, where L represents the location of the feature, σ and θ denote the point's characteristic scale and dominant orientation, respectively, and d is a 128-vector invariant feature descriptor. Then, feature points are matched using the minimum Euclidean distance method between their descriptors. To ensure correct matching, Lowe [25] suggested another matching strategy called NNDR (nearest neighbor distance ratio), which means that the ratio of the nearest to the second nearest neighbor can be applied to get correct matches. It can be described as follows:

$$NN_1/NN_2 \leq d_{ratio} \tag{6}$$

where, NN_1 means the Euclidean distance from the nearest vector in sensed image feature map to the vector in reference image feature map and NN_2 is the Euclidean distance from the second nearest vector in sensed image feature map to the vector in reference image feature map. When the distance ratio is lower than d_{ratio}, it can be seen as a pair of correct matches.

However, SIFT does not perform well when geometric distortion between images is severe. In order to improve the performance in this situation, ASIFT [26] is proposed, which simulates the rotation of camera around optical axis. In their method, image affine transformation is applied to model the changes of viewpoints, which can be expressed as:

$$u(x, y) \rightarrow u(ax + by + e, cx + dy + f) \tag{7}$$

The processing steps of ASIFT can be summarized as follows: first, a dense set of rotation transformation is applied to both images A and B; then a series of tilt transformation is applied to all rotated images; at last, SIFT is performed to match all pairs of the simulated images.

3.3. Integration of 2D Vector BFPZs and Surveillance Video Images

In the proposed method, the mapping between 2D Vector BFPZs and PTZ surveillance video is constructed in a semi-automatic way. To solve the homography matrix between BFPZs data and the preset frame of PTZ camera, we provide an interactive tool to select control points. Once an adequate number of points are matched, Levenberg–Marquardt iterative optimization [27,28], is applied to calculate the homography matrix.

Assuming that 2D Vector BFPZs is a set of polygons $\{P_1, P_2, \ldots, P_n\}$, for a given polygon $P_i\{(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), \ldots, (X_{is}, Y_{is})\}, i = 1, 2, \ldots, n$, and the coordinates of its vertex are $V\left(X_{ij}, Y_{ij}\right), j = 1, 2, \ldots, s$. Then this vertex can be mapped into an image coordinate system, and its coordinates are $V'(x_{ij}', y_{ij}')$, which can be computed as follows.

$$\begin{cases} x_{ij}' = \frac{A_0 X_{ij} + B_0 Y_{ij} + C_0}{G_0 X_{ij} + H_0 Y_{ij} + 1} \\ y_{ij}' = \frac{D_0 X_{ij} + E_0 Y_{ij} + F_0}{G_0 X_{ij} + H_0 Y_{ij} + 1} \end{cases}, \tag{8}$$

where, A_0, B_0, C_0, D_0, E_0, F_0, G_0, H_0, are parameters of $\mathbf{H}_0$.

The precision, number, and spatial distribution of the selected points play a vital role in the process of calculating the homography matrix, which directly affects the accuracy of the estimated matrix. It is better to choose relative permanent points, such as road intersection, as shown in Figure 3.

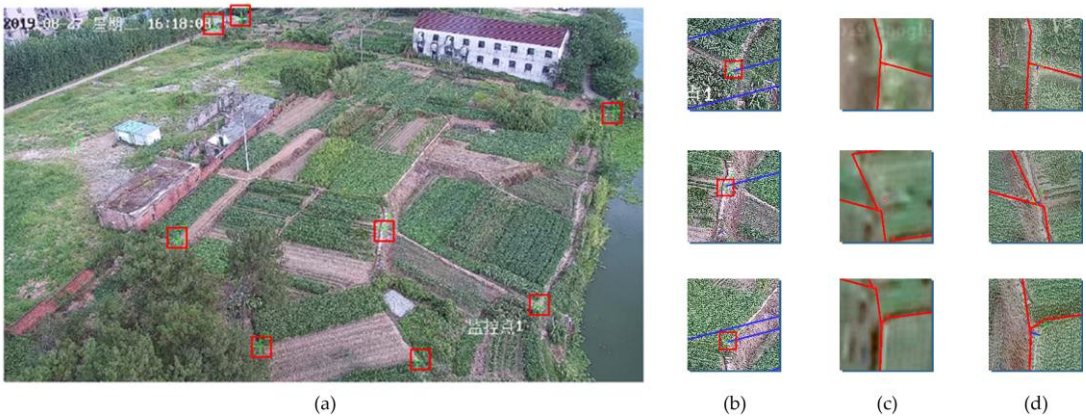


Figure 3. (**a**) The distribution of selected control points. (**b**) Local video image of selected features. (**c**) Local remote sensing image of selected features with 2D vector basic farmland protection zones (BFPZs) overlaid. (**d**) Local video image of selected features with 2D vector BFPZs overlaid.

When the PTZ camera is paned, tilted, or zoomed, the homography matrix between GIS data and video frames can be calculated as follows. First, ASIFT matching method is used to detect and match features from video sequences. Then, random sample consensus (RANSAC) [29] is applied to estimate the homography matrix between video frames, considering that the obtained corresponding features have mismatches even using the best automatic matching method. RANSAC estimates the parameters of homography model from a set of observed data containing outliers in an iterative way. Finally, the mapping between 2D Vector BFPZs and any video frame can be calculated.

Specifically, the locations of 2D Vector BFPZs in each single video frame coordinates system can be computed by two strategies. What counts is the selection of reference video frame. If the preset video frame is chosen as the reference image, ASIFT and RANSAC methods are applied to estimate the mapping matrix between video frames and preset image. As shown in Figure 4, assuming that the estimated transformation matrix between reference frame and frame I_t is $\mathbf{H}_t$, and any vertex V_t' of the BFPZs polygons in video frame I_t can be calculated using its corresponding vertex V' of the BFPZs polygons in reference frame as the following equation.

$$V_t' = \mathbf{H}_t V', \tag{9}$$

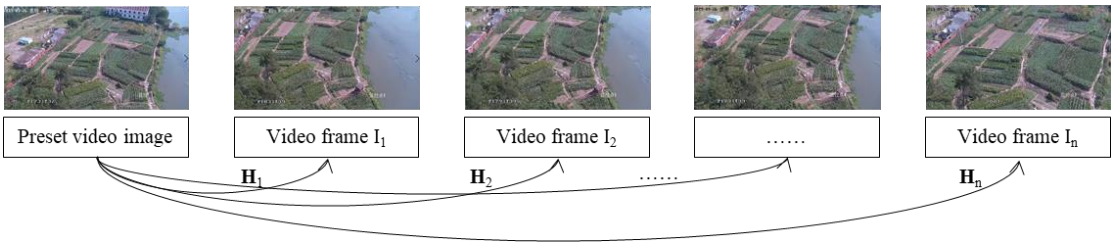


Figure 4. Transformations between video frames and the preset image.

Image matching is performed between adjacent video frame, as shown in Figure 5. In this situation, vertex V_t' of the BFPZs polygons in video frame I_t can be computed as Equation (10).

$$V_t' = \mathbf{H}_t \cdots \mathbf{H}_2 \mathbf{H}_1 V', \tag{10}$$

where $\mathbf{H}_i$ is the transformation matrix between video frame I_{i-1} and I_i.

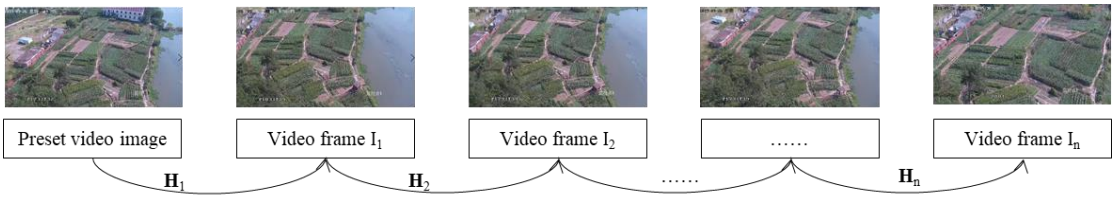


Figure 5. Transformations between adjacent video frames.

4. Framework of Cultivated Land Surveillance System

On the basis of the integration method of 2D vector BFPZs and real-time PTZ surveillance video, we design and implement a prototype of monitoring system. The system can collect, manage geospatial and video data, and perform overall display and analysis. Through video surveillance of cultivated land, supervisors can efficiently improve the supervision work. As shown in Figure 6, the integration framework mainly contains three layers: presentation layer, middle layer, and data layer.

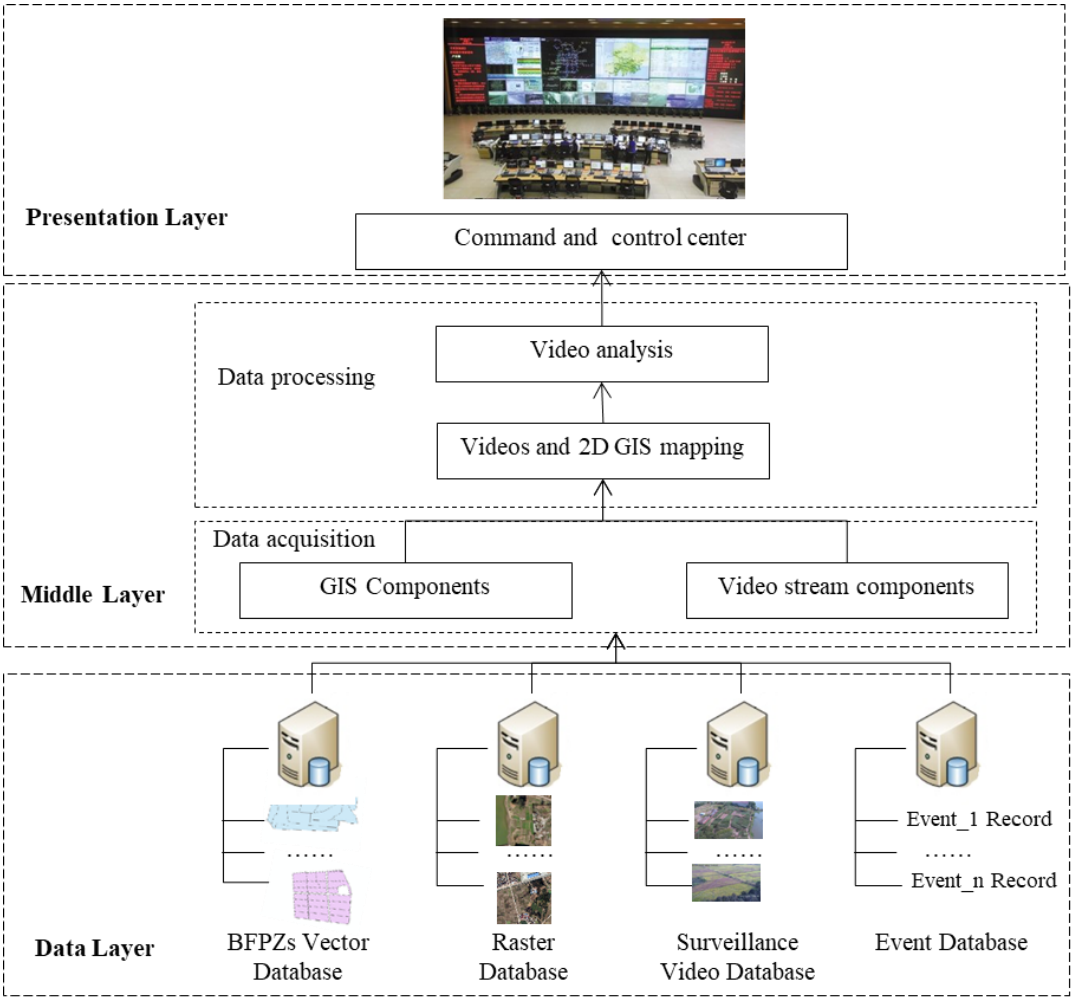


Figure 6. Framework diagram of the proposed comprehensive monitoring method for cultivated land monitoring.

1. **Data layer**: The data layer is mainly used to store and manage geospatial and surveillance video data. In the proposed system, geospatial data contains BFPZs presented in 2D vector format and high-resolution ortho-images presented in raster format. The former is used to define ROI

regions in real-time video, and the latter is used to establish the mapping relationship between geospatial space and image pixel coordinates.

2. **Middle layer**: The middle layer has functions of data acquisition, processing, and analysis. As geospatial data and video data are unstructured data, they are independent of each other. In this layer, GIS components are used for fetching geospatial data, and video stream components are used for reading real-time video data. 2D mapping between BFPZs and video images is the core component, and BFPZs data can be projected into real-time video to generate ROI area. For the specific area, video analysis can be performed.
3. **Presentation layer**: In the presentation layer, it provides interactive integration of 2D BFPZs data and surveillance video as well as real-time alert for command and control center.

The software structure of the cultivated land monitoring system consists of five parts: video retrieval, video control, image registration, image matching, and video analysis. Video retrieval is mainly used to load real-time video using HTTP protocol and display the current video frame. It can also visualize 2D BFPZs overlaid on video frames. Video control allows the user to pan, tilt, and zoom the PTZ camera and fetch the current, previous, or next video frame. Image registration provides an interactive tool for constructing the mapping matrix between 2D vector BFPZs and the preset video image by selecting correspondences between image space and geospatial space and solving the matrix parameters. Then, 2D vector BFPZs is projected into the preset video image. Image matching provides an automatic way that can use the multi-view image matching method ASIFT to construct the mapping between surveillance video images. Correspondences between video images can be detected and matched without human involvement. The projected 2D vector BFPZs in the preset video image can be mapped into real-time video images with different matching strategies. At last, 2D vector BFPZs polygons can be set as the defense area for intelligent video surveillance, which can efficiently extract useful information from the huge number of videos by automatically detecting, tracking, and recognizing objects and analyzing activities [14].

5. Experiments and Results

Details of experiments on the alignment of 2D vector data and video images are presented in this section. We tested our system in cultivated land with relatively smooth landscape and compared the alignment results with man-made ground truth.

5.1. Datasets and Evaluation Criteria

In order to visually and quantitatively evaluate the proposed method, two intelligent surveillance cameras (PTZ) were deployed on China Mobile's signal tower in Dongyang City, Zhejiang Province, China. One is located in the Woodcarving town, and the other one is located in Wangfeng community, Jiangbei Street. Meanwhile, two video datasets were recorded and collected from the two surveillance cameras, detailed information can be seen in Table 1. Dataset MDXZ contains the scenes of cultivated land in Woodcarving town, while JBWF contains the scenes of Wangfeng community. These two datasets were created by rotating the cameras vertically and horizontally, as well as changing the zoom level. As shown in Figure 7, cultivated landscape in the two places is relatively smooth.

Table 1. The experimental datasets.

Video Datasets	Size	Brief Description
MDXZ	1920 × 1080	Large parcels of arable land; large geometric distortion between video images
JBWF	1920 × 1080	Small size of arable land; little geometric distortion between video images

(a) (b)

Figure 7. Scenes of the two datasets. (**a**) MDXZ and (**b**) JBWF.

The mapping between 2D vector BFPZs and video frames is taken as an issue of image registration. Therefore, root of mean square error (RMSE) is used to qualify the accuracy of alignment results, which measures the deviation between the observed value and ground truth. It can be calculated as follows:

$$RMSE = \sqrt{\frac{(x - x')^2 + (y - y')^2}{M}}, \tag{11}$$

where $(x,\ y)$ is the coordinates of the selected M points in transformed 2D GIS data, and $(x',\ y')$ demotes the ground true values.

Since we lack ground truth values, it is hard to determine the projection error of the estimated mapping relationship. In addition, pixel coordinates system between video frames is independent. To cope with the above challenge, we manually constructed ground truth for each video frame as follows:

(1) Select 20 points in 2D vector BFPZs and record their projected coordinates in the image pixel coordinate system;
(2) For each point in (1), find its corresponding point by visual inspection and record its coordinates in the current image pixel system.

Then, for each video frame, the projection error can be calculated by using RMSE. Note that some principles should be obeyed when selecting points. For example, the selected points should be well-distributed.

5.2. Implementations and Experimental Results

The cultivated land monitoring system was implemented as a Visual C# project in Microsoft Visual Studio 2017 and works in a client/server environment. This application is mainly designed for PTZ network (i.e., IP) cameras. It is also suitable for fixed cameras. As Figure 8 shows, the user can retrieval real-time video using the HTTP protocol. The URL, username, and password are needed to access the camera API to retrieve the current video frame.

The user interface is composed of two interconnected components. The first involves the registration of 2D BFPZs data from the geospatial coordinates system to video image pixel coordinate system, which corresponded to the function of image registration. The second involves registering 2D BFPZs data in different video image space, which corresponded to the function of image matching.

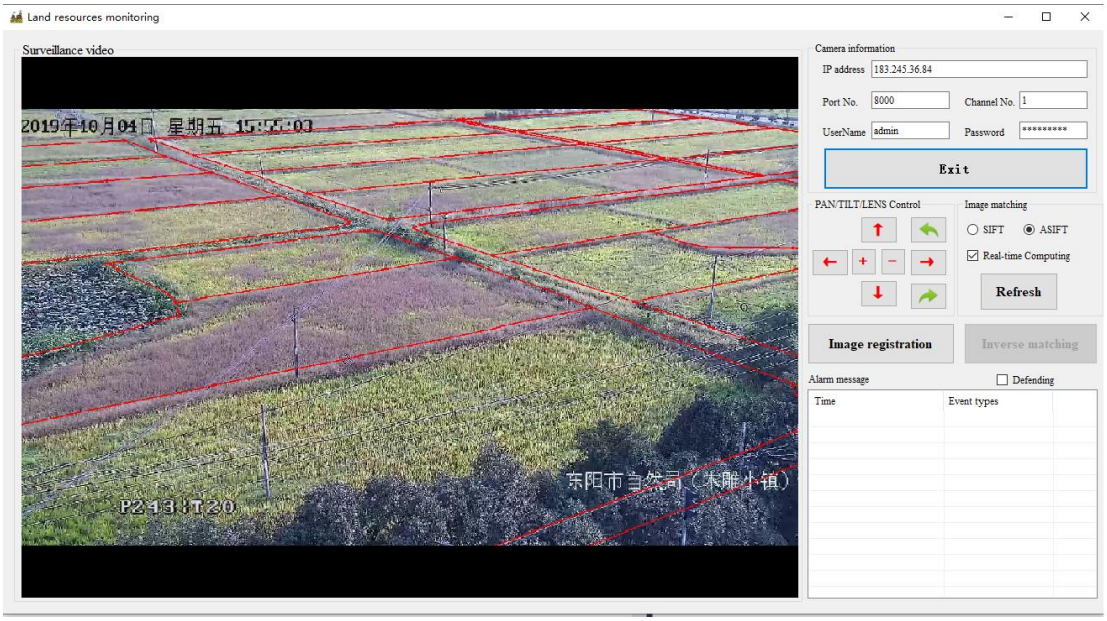
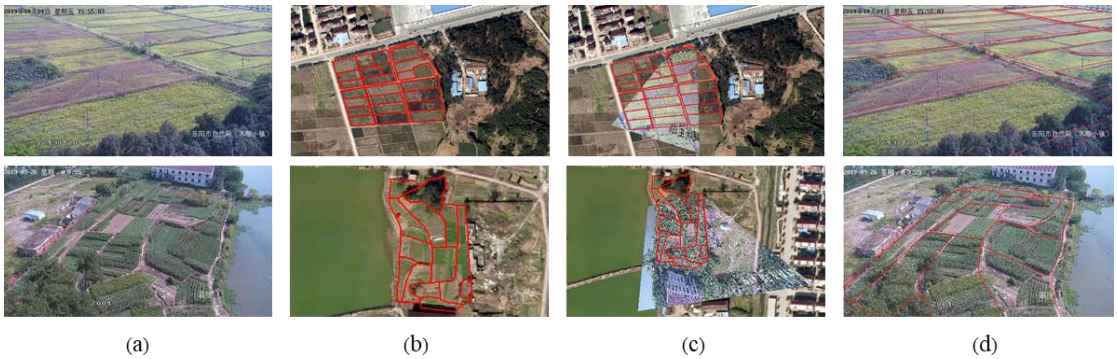


Figure 8. The user interface of video-based surveillance system for cultivated land.

To visually check the suitability of the proposed method, several examples are taken in Figures 8 and 9. Though the intention of this paper is to register 2D BFPZs into real-time video to determine the defense boundary of cultivated land, in fact, this is a reversible process, which means that the real-time video image is also mapped into the geospatial space. This mapping can satisfy the needs of spatial analysis, such as the measurement of cultivated land. Figure 9 shows the cross-mapping results of 2D BFPZs (polygons in red) and the preset video image, while Figure 10 shows the matching results between video images and results of projected 2D BFPZs in different video image space using ASIFT and RANSAC method. From these figures, we can see that the proposed method can successfully build the transformation between 2D BFPZs and video image, and accurately define the boundaries of the protected cultivated land in the video image.



(a) (b) (c) (d)

Figure 9. The cross-mapping between 2D BFPZs and the preset video frame of MDXZ (upper row) and JBWF (lower row) dataset. (**a**) The preset video frame. (**b**) The corresponding ortho-image overlaid with 2D vector BFPZs. (**c**) The preset video frame projected into the geospatial space. (**d**) The 2D vector BFPZs projected into the image space.

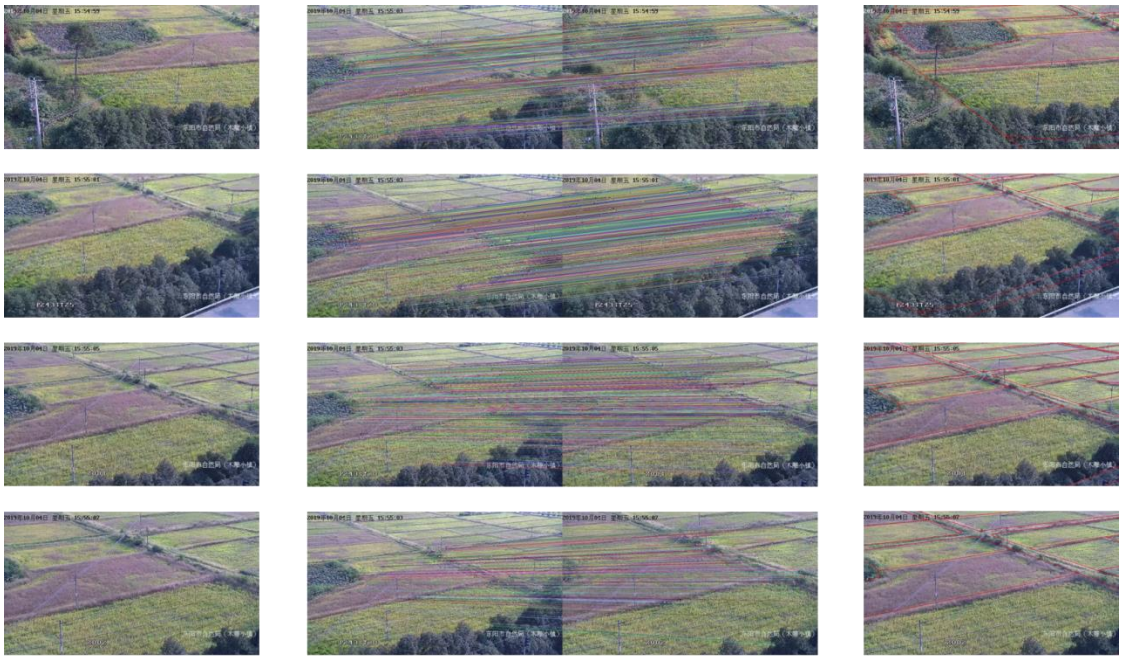


Figure 10. 2D BFPZs projected into different image space by using the proposed method in the MDXZ dataset. (**a**) Part frames of video. (**b**) The matching results, in which corresponding points are connected with a straight line. (**c**) Video frames overlaid with the projected BFPZs.

5.3. Accuracy Analysis

The examples presented in Figures 9 and 10 show the feasible and achievable accuracy of the solution for projecting 2D vector BFPZs into real-time video frames. In this section, we mainly focus on the evaluation of the proposed method in a quantitative way and comprehensively analyze the influence of the selected reference image on the alignment results.

In this experiment, image video frames named MDXZ05 and JBWF05 in the two datasets were chosen as the preset image, respectively. We manually selected 6–8 control points to map the 2D BFPZs from the geospatial coordinate system to the preset image pixel system. Then two matching strategies (Strategy-A: adjacent image is set as the reference image; Strategy-S: preset image is set as the reference image;) are applied to automatically detect and match control points to estimate the mapping matrix. The relationship between images can be depicted in Figures 11 and 12, taking MDXZ dataset as an example.

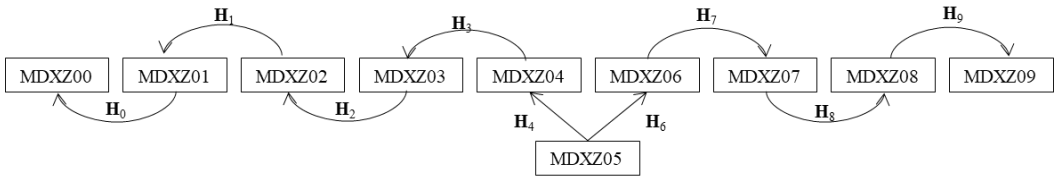


Figure 11. The relationship between video frames when using Strategy-A.

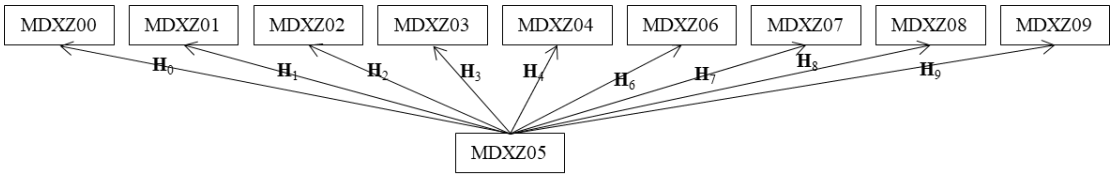


Figure 12. The relationship between video frames when using Strategy-S.

Quantitative comparisons of different matching strategies and two datasets using the RMSE are reported in Tables 2 and 3, respectively. RMSE-A represents the results of Strategy-A, and RMSE-S represents the results of Strategy-S.

Table 2. Experimental results on dataset MDXZ. Quantitative comparisons on different matching strategies using the root of mean square error (RMSE).

Video Frames	RMSE-A (Pixels)	RMSE-S (Pixels)
MDXZ00	15.08573	12.10606
MDXZ01	10.59250	8.99473
MDXZ02	7.78827	7.192565
MDXZ03	9.35951	9.39466
MDXZ04	9.63640	9.63640
MDXZ05	*	*
MDXZ06	6.53356	6.53356
MDXZ07	13.50897	13.43115
MDXZ08	15.93654	14.57003
MDXZ09	23.46391	22.34980
Average	12.43393	11.57877

* denotes no data.

Table 3. Experimental results on dataset JBWF. Quantitative comparisons on different matching strategies using the RMSE.

Video Frames	RMSE-A (Pixels)	RMSE-S (Pixels)
JBWF00	16.3613	16.10768
JBWF01	9.32134	9.26386
JBWF02	6.08594	6.04870
JBWF03	7.23469	7.21215
JBWF04	11.04419	11.04419
JBWF05	*	*
JBWF06	8.89209	8.89209
JBWF07	12.76803	12.76073
JBWF08	12.92818	12.83391
JBWF09	17.74984	17.50123
Average	11.37618	11.29606

* denotes no data.

From Tables 2 and 3, we can see that the average projection error of the two datasets is within 13 pixels. As the surveillance camera get images in a side view, the GSD (ground sample distance) varies with the distance from the object to the camera. With reference to the GSD of TDOMs (true digital orthophoto maps) generated from low-altitude UAVs images [30], it is believed that the proposed integration method can achieve meter-accuracy and meet the requirements of real-time cultivated land monitoring in the experimental area. Meanwhile, one can observe that the average image registration accuracy of dataset MDXZ is higher than that of dataset JBWF, which can also reflect that the geometric difference between video images in MDXZ is greater than that in JBWF.

By comparing the experimental results of video frames in MDXZ dataset, a clear trend stands out: both RMSE-A and RMSE-S increase when the matching process varies from MDXZ05 to MDXZ00 and

from MDXZ05 to MDXZ09, in which the geometric distortion becomes larger when the PTZ camera is panned, tilted, and zoomed. Furthermore, it can be found that Strategy-S can obtain higher accuracy than Strategy-A. This is because the matching error will accumulate as the matching times increase. An interesting discovery is that the best image registration results may not be obtained between the nearest image, but may be the second or third nearest image, such as the results of MDXZ02, MDXZ03, and MDXZ04 in Table 2. A similar conclusion can be reached in Table 3.

6. Discussions

For a GIS-based video monitoring system, the prerequisite is to align the GIS data with video images. In order to accomplish this alignment, it is necessary to establish a mapping relationship between the video frame and GIS data. In this study, we propose a method to register 2D vector data of the protected cultivated land onto surveillance video images. Experimental results in cultivated land with relatively flat landscape suggest that the proposed method can accurately project 2D vector BFPZs onto real-time video frames and obtain meter-level positioning accuracy to meet the requirements of cultivated land monitoring.

Currently, there have been some studies on GIS-based video monitoring systems. The GeoScopeAVS system designed by Milosavljevic et al. [16] integrated GIS and video surveillance for real-time retrieval of information about viewed geospatial objects. In their system, 3D GIS and camera views were aligned so that it is possible to extract the visualization of appropriate 3D virtual objects and place them over the video. However, this method may not be applicable to cultivated land, because 3D models of geographic features and accurate DEM of the area are essential. As we all know, people pay more attention to the modeling of urban environment, especially buildings, but the modeling of rural environment is largely ignored. Moreover, not all objects in the real world can be completely modelled, such as vehicles, people, trees, and street lights [15]. Zhao et al. [17] designed a GIS-based video monitoring system for forest insect defoliation and discoloration. Though 2D GIS and 3D GIS are simultaneously integrated in their system, it is not a real fusion because these data are geographically linked and separately displayed in their own window. Nonetheless, it is difficult for non-experts to interpret these data as they are not presented in the same layer. Milosavljevic et al. [19] proposed a method to georeference video images, which relied on matching video frame coordinates of certain point features with their 3D geographic locations. Overall, the above-mentioned methods have one thing in common: all of them require internal and external parameters of the camera and high-precision 3D geospatial data, which are difficult to be accurately obtained for the case of cultivated land and cause failure. Zhang et al. [18] presented a GIS-based prototype system for city safety management. They described a semi-automatic cross-mapping method based on 2D GIS and some constraints. However, the case of the PTZ camera was not considered in their system, which made them unsuitable for the monitoring of cultivated land. In comparison to them, our solution is more feasible in monitoring cultivated land. On the one hand, it is fast and easy to be implemented. Neither of the camera parameters and 3D model of objects are needed to integrate 2D GIS data and PTZ video images. Once a number of control points are identified from 2D GIS data and the selected reference video image, the alignment of 2D GIS data and PTZ video frames can be realized by automatic feature matching method. On the other hand, even non-experts and inexperienced persons can intuitively see the boundary of protected farmland from the surveillance video and identify destruction of cultivated land.

There are some factors that may influence the performance of our system: the selection of reference images as well as the amplitude and speed of video camera movement may change the overlap between images and have an impact on feature matching. If overlap between images is too small, it is difficult to obtain uniformly distributed corresponding features to estimate the geometric transformation matrix. In addition, a different matching strategy may affect the accuracy of video images and GIS data registration results. As described in Section 5.2, a reasonable matching strategy can reduce the propagation of matching error. Comprehensively, it is better to choose Strategy-S instead of Strategy-A

for the integration of 2D vector BFPZs and video images. One reason is the error propagation in Strategy-A, and the other reason is that it is difficult to determine the interval between adjacent images. The major shortcoming of Strategy-S is that the movement of the PTZ camera can be restricted. If the overlap between the preset frame and the current frame is too small, this strategy may fail. Furthermore, the proposed method is not suitable for areas of complex terrain, such as mountainous areas. In such areas, the geometric mapping relationship between video image plane and 2D GIS data cannot be simply modelled by homography transformation. It may fail in alignment video images with GIS data, and prevents the following application.

Based on the alignment results, we can easily identify the occurrence of farmland destruction by visually inspecting the image content covering the 2D vector area. In China, the conversion of permanent farmland in BFPZs to other uses is forbidden, while the use nature of non-permanent agricultural land can be changed, for example, to construction land, with the approval of relevant departments. Since the two kinds of farmland are very similar and indistinguishable in video images, it is necessary to label the protected area on the video image. The proposed method provides a solution to solve the problem of labelling basic farmland in PTZ videos. Polygons in the 2D GIS data define the boundary of protected cultivated land. If image content delimited by polygons is recognized as construction land, it can be recognized as an illegal change.

However, little effort has been made to automatically identify destruction of farmland or unauthorized change of farmland use from the video images. In current literature, there is almost no research and report on this issue. Moreover, this issue is a complicated problem that requires technologies in various subjects and fields, such as computer vision, intelligent video analysis, image processing, and data mining. In the future, we will try to improve the robustness of our system and further study on automatic detection of illegal farmland occupation from real-time PTZ video images delineated by 2D GIS vector data.

7. Conclusions

The objective of this paper is to integrate 2D GIS and real-time PTZ surveillance, and implement a prototype of a cultivated land monitoring system based on the proposed integration method. This integration can assist users in identifying the illegal occupation of the protected cultivated land as early as possible. Since the permanent cultivated land and non-permanent agriculture land are very similar and indistinguishable in video images, 2D GIS data can provide the boundaries and other semantic information of the protected farmland. For the integration process, the alignment of GIS data and video frames is necessary. In the proposed method, 2D GIS vector data is projected onto the video image by the transformation matrix, which is estimated from the extracted pairs of corresponding features. Compared with the existing methods or systems, the proposed integration method has the following advantages: (1) it is fast and easy to be implemented. Neither of the camera parameters and 3D model of objects or terrain are needed to integrate 2D GIS data and PTZ video images. (2) It is easier for non-experts and inexperienced persons to identify illegal occupation of the protected land from the real-time video. In addition, there are some other issues worthy of consideration. For example, how to automatically identify destruction of farmland or unauthorized change of farmland use from the video images with the aid of projected 2D GIS data. This will be our future studies.

Author Contributions: Conceptualization, Zhenfeng Shao and Congmin Li; Methodology, Zhenfeng Shao and Congmin Li; Validation, Congmin Li, Lin Ding and Lei Zhang; Formal Analysis, Zhenfeng Shaoand Congmin Li; Data Curation, Congmin Li, Lei Zhang and Lin Ding; Writing—Original Draft Preparation, Zhenfeng Shao and Congmin Li; Writing—Review and Editing, Congmin Li, Deren Li, Orhan Altan and Lei Zhang; Supervision, Congmin Li; Funding Acquisition, Zhenfeng Shao and Lei Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China (2018YFB0505401), the Research Project from the Ministry of Natural Resources of China under Grant 4201-240100123, the National Natural Science Foundation of China under Grants 41771452, 41771454, 41890820, and 41901340, the Natural Science Fund of Hubei Province in China under Grant 2018CFA007, the Open Fund

of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources under Grant KF-2019-04-048.

Acknowledgments: The authors are sincerely grateful to the editors as well as the anonymous reviewers for their valuable suggestions and comments that helped us improve this paper significantly.

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

References

1. *China Statistical Yearbook 2019*; National Bureau of Statistics of the People's Republic of China: Beijing, China, 2019. Available online: http://www.stats.gov.cn/tjsj/ndsj/2019/indexch.htm (accessed on 13 May 2020).
2. Yuan, Y.; Lin, L.; Chen, J.B.; Sahli, H.C.; Chen, Y.X.; Wang, C.Y.; Wu, B. A New Framework for Modelling and Monitoring the Conversion of Cultivated Land to Built-up Land Based on a Hierarchical Hidden Semi-Markov Model Using Satellite Image Time Series. *Remote Sens.* **2019**, *11*, 210. [CrossRef]
3. Wang, Y.; Gao, J.X.; Zou, C.X.; Xu, D.L.; Wang, L.X.; Jin, Y.; Wu, D.; Lin, N.F.; Xu, M.J. Identifying ecologically valuable and sensitive areas: A case study analysis from China. *J. Nat. Conserv.* **2017**, *40*, 49–63. [CrossRef]
4. Cheng, Q.W.; Jiang, P.H.; Cai, L.Y.; Shan, J.X.; Zhang, Y.Q.; Wang, L.Y.; Li, M.C.; Li, F.X.; Zhu, A.X.; Chen, D. Delineation of a permanent basic farmland protection area around a city centre: Case study of Changzhou City, China. *Land Use Pol.* **2017**, *60*, 73–89. [CrossRef]
5. Coughlin, R.E. FORMULATING AND EVALUATING AGRICULTURAL ZONING PROGRAMS. *J. Am. Plan. Assoc.* **1991**, *57*, 183–192. [CrossRef]
6. Dou, P.; Chen, Y.B. Dynamic monitoring of land-use/land-cover change and urban expansion in Shenzhen using Landsat imagery from 1988 to 2015. *Int. J. Remote Sens.* **2017**, *38*, 5388–5407. [CrossRef]
7. Hussain, M.; Chen, D.M.; Cheng, A.; Wei, H.; Stanley, D. Change detection from remotely sensed images: From pixel-based to object-based approaches. *Isprs J. Photogramm. Remote Sens.* **2013**, *80*, 91–106. [CrossRef]
8. Wei, Z.Q.; Han, Y.F.; Li, M.Y.; Yang, K.; Yang, Y.; Luo, Y.; Ong, S.H. A Small UAV Based Multi-Temporal Image Registration for Dynamic Agricultural Terrace Monitoring. *Remote Sens.* **2017**, *9*, 904. [CrossRef]
9. Song, F.; Dan, T.T.; Yu, R.; Yang, K.; Yang, Y.; Chen, W.Y.; Gao, X.Y.; Ong, S.H. Small UAV-based multi-temporal change detection for monitoring cultivated land cover changes in mountainous terrain. *Remote Sens. Lett.* **2019**, *10*, 573–582. [CrossRef]
10. Song, F.; Li, M.Y.; Yang, Y.; Yang, K.; Gao, X.Y.; Dan, T.T. Small UAV based multi-viewpoint image registration for monitoring cultivated land changes in mountainous terrain. *Int. J. Remote Sens.* **2018**, *39*, 7201–7224. [CrossRef]
11. Ma, L.; Cheng, L.; Han, W.Q.; Zhong, L.S.; Li, M.C. Cultivated land information extraction from high-resolution unmanned aerial vehicle imagery data. *J. Appl. Remote Sens.* **2014**, *8*, 25. [CrossRef]
12. Ai, M.Y.; Hu, Q.W.; Li, J.Y.; Wang, M.; Yuan, H.; Wang, S.H. A Robust Photogrammetric Processing Method of Low-Altitude UAV Images. *Remote Sens.* **2015**, *7*, 2302–2333. [CrossRef]
13. Zhang, Y.J.; Xiong, J.X.; Hao, L.J. Photogrammetric processing of low-altitude images acquired by unpiloted aerial vehicles. *Photogramm. Rec.* **2011**, *26*, 190–211. [CrossRef]
14. Wang, X.G. Intelligent multi-camera video surveillance: A review. *Pattern Recognit. Lett.* **2013**, *34*, 3–19. [CrossRef]
15. Milosavljevic, A.; Rancic, D.; Dimitrijevic, A.; Predic, B.; Mihajlovic, V. Integration of GIS and video surveillance. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 2089–2107. [CrossRef]
16. Milosavljevic, A.; Dimitrijevic, A.; Rancic, D. GIS-augmented video surveillance. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1415–1433. [CrossRef]
17. Zhao, F.F.; Wang, Y.F.; Qiao, Y.Y. Framework for video-based monitoring of forest insect defoliation and discoloration. *J. Appl. Remote Sens.* **2015**, *9*, 15. [CrossRef]
18. Zhang, X.; Liu, X.; Song, H. Video surveillance GIS: A novel application. In Proceedings of the IEEE 21st International Conference on Geoinformatics, Kaifeng, China, 20–22 June 2013; pp. 1–4.
19. Milosavljevic, A.; Rancic, D.; Dimitrijevic, A.; Predic, B.; Mihajlovic, V. A Method for Estimating Surveillance Video Georeferences. *Isprs Int. J. Geo-Inf.* **2017**, *6*, 211. [CrossRef]
20. Xie, Y.J.; Wang, M.Z.; Liu, X.J.; Wu, Y.G. Integration of GIS and Moving Objects in Surveillance Video. *Isprs. Int. J. Geo-Inf.* **2017**, *6*, 94. [CrossRef]

21. Reulke, R.; Bauer, S.; Döring, T.; Meysel, F. Traffic Surveillance using Multi-Camera Detection and Multi-Target Tracking. In Proceedings of the Image and Vision Computing, Hamilton, New Zealand, 5–7 December 2007; pp. 175–180.
22. Tian, Y.; Chen, C.; Shah, M. Cross-View Image Matching for Geo-Localization in Urban Environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1998–2006.
23. Regmi, K.; Shah, M. Bridging the Domain Gap for Ground-to-Aerial Image Matching. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 470–479.
24. Lin, T.; Belongie, S.; Hays, J. Cross-View Image Geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 891–898.
25. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
26. Morel, J.M.; Yu, G.S. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *Siam. J. Imaging Sci.* **2009**, *2*, 438–469. [CrossRef]
27. Marquardt, D.W. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431–441. [CrossRef]
28. Levenberg, K. A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Q. Appl. Math.* **1944**, *2*, 164–168. [CrossRef]
29. Fischler, M.A.; Bolles, R.C. Random sample consensus—A paradigm for model-fitting with applications to image-analysis and automated cartography. *Commun. Acm.* **1981**, *24*, 381–395. [CrossRef]
30. Liu, Y.; Zheng, X.; Ai, G.; Zhang, Y.; Zuo, Y. Generating a High-Precision True Digital Orthophoto Map Based on UAV Images. *Isprs. Int. J. Geo-Inf.* **2018**, *7*, 333. [CrossRef]