International Journal of Geo-Information

MDPI

Article

Integration of Multi-Camera Video Moving Objects and GIS

Yujia Xie [1], Meizhen Wang [2,3,4,*], Xuejun Liu [2,3,4], Bo Mao [1] and Feiyue Wang [1]

1 Key College of Information Engineering, Nanjing University of Finance & Economics, Nanjing 210023, China; 9120181003@nufe.edu.cn (Y.X.); bo.mao@nufe.edu.cn (B.M.); 1120180759@stu.nufe.edu.cn (F.W.)
2 Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China; liuxuejun@njnu.edu.cn
3 State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China
4 Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
* Correspondence: wangmeizhen@njnu.edu.cn

Received: 10 October 2019; Accepted: 2 December 2019; Published: 7 December 2019

Abstract: This work discusses the integration of multi-camera video moving objects (MCVO) and GIS. This integration was motivated by the characteristics of multi-camera videos distributed in the urban environment, namely, large data volume, sparse distribution and complex spatial–temporal correlation of MCVO, thereby resulting in low efficiency of manual browsing and retrieval of videos. To address the aforementioned drawbacks, on the basis of multi-camera video moving object extraction, this paper first analyzed the characteristics of different video-GIS Information fusion methods and investigated the integrated data organization of MCVO by constructing a spatial–temporal pipeline among different cameras. Then, the conceptual integration model of MCVO and GIS was proposed on the basis of spatial mapping, and the GIS-MCVO prototype system was constructed in this study. Finally, this study analyzed the applications and potential benefits of the GIS-MCVO system, including a GIS-based user interface on video moving object expression in the virtual geographic scene, video compression storage, blind zone trajectory deduction, retrieval of MCVO, and video synopsis. Examples have shown that the integration of MCVO and GIS can improve the efficiency of expressing video information, achieve the compression of video data, rapidly assisting the user in browsing video objects from multiple cameras.

Keywords: geovisualization; GIS; video-GIS; surveillance video; virtual reality

1. Introduction

At present, billions of surveillance video cameras have been deployed worldwide. The images acquired from these cameras are widely used in security, transportation, environmental detection, and other fields to monitor real-time changes in geographic scenes 24 hours per day. In the actual monitoring process, displaying multiple video images in a grid interface cannot effectively express the spatial relationship among different camera images in the urban environment (Figure 1). Furthermore, monitoring tasks, such as spatial–temporal behavior analysis, video scene simulation, and regional condition, cannot be effectively completed by only relying on image data. To solve the aforementioned problems, monitoring images should be introduced into the geographic information systems (GIS) to construct a video-GIS (V-GIS) surveillance system. V-GIS is a geographical environment perception and analysis platform that integrates a traditional video analysis system within GIS. Under the unified geographic reference, a geospatial data service supports surveillance image intelligence analysis and

realizes image–scene integrated modeling [1], video data spatialization [2], video data management analysis [3], virtual reality (VR) fusion expression [4], and other related functions. Unlike other GIS data, videos have large data volume, sparse distribution of video moving objects (like pedestrians and vehicles), and complex spatial–temporal correlation of video moving objects from different cameras. These characteristics result in the low efficiency of manual browsing and video retrieval and an insufficient ability to analyze a videos' geospatial correlation. As a result, traditional GIS data research methods cannot effectively analyze video data, and a specific research method for V-GIS must be determined.

Figure 1. Video images from multiple cameras expressed on a grid interface.

Currently, studies on V-GIS involves a series of investigations on the specificity of video data, including geographic video semantic expression [5], structured processing on geo-video [6], video integration and GIS [1]. However, most existing research focuses on the integration of video image and GIS, thereby ignoring the integration of video moving object and GIS. Video moving objects are the main focus of users, whereas traditional manual retrieval and analysis on moving objects from massive surveillance video requires considerable computational resources and time and generates many misdetections and misjudgments. In recent years, new algorithms, such as mask region-convolutional neural network (R-CNN) [7] and "you only look once" (YOLO.v3) [8], which improves the efficiency and accuracy of video moving object detection and rapid recognition and make the information fusion between video moving objects and GIS accurate and feasible has been presented. A new fusion type on the integration of video moving object and GIS should be determined to enhance the effectivity of the retrieval and analysis in V-GIS.

In this work, we considered the integration of multi-camera video moving objects (MCVO) and GIS. This integration was essentially an augmented VR technique applied in GIS [9]. The integration of single camera video moving objects and GIS has been discussed before [10]. Practically, V-GIS is oriented toward multi-camera video information processing in a surveillance network, in which a single video moving object may appear in multiple camera fields of view, and its trajectory and semantic behavior have complex spatial–temporal associations related to multiple camera shots [11]. Therefore, a system that can effectively organize the video moving objects associated with multiple cameras should be proposed and constructed to perform comprehensive geospatial processing and analysis.

This study attempted to construct an integrated MCVO and GIS system. This integration could achieve not only the video moving object information extraction and spatially-correlated visualization but also the MCVO's spatial–temporal associated analysis, which effectively assists users in understanding multi-camera videos. The main contributions of this paper are presented as follows: 1. The conceptual fusion model between MCVO and GIS is proposed by comparing the characteristics of different V-GIS fusion methods on the basis of the specificity of MCVO organization. 2. The GIS-MCVO prototype system is proposed by describing the architecture and function design of the system. 3. The unique functions and benefits of GIS-MCVO system, including GIS-based user interface on video moving object expression, video compression storage, blind zone trajectory deduction, retrieval of multiple camera video objects, and video synopsis, which cannot be easily achieved in the traditional V-GIS integrated system, are analyzed.

This study only investigated the video data obtained by camera with a fixed position and attitude. The organizational structure of this work is presented as follows. Section 2 presents an overview of the related work. Section 3 discusses the fusion modes and the organization of MCVO integrated with 3D GIS. Section 4 describes the architecture of the GIS-MCVO prototype system and analyzes its main features and key technologies. Section 5 lists the related applications and the potential advantages of the GIS-MCVO system. Section 6 summarizes and concludes the study.

2. Related Work

A MCVO and GIS integrated monitoring system were developed on the basis of video moving object detection and V-GIS integrated data organization technology to express video information in the virtual geographic environment. This section describes the related research on intelligent surveillance videos, video moving object extraction, V-GIS integrated data organization, and V-GIS fusion expression.

Although the surveillance video technology has been partially used in the last century [12,13], video surveillance systems that consist of closed-circuit television and analog signals have poor range monitoring capabilities, information storage, and comprehensive analysis. Since the "9/11" incident in 2001, the demand for enhancing video surveillance has greatly increased along with the amount of surveillance image data, thereby causing great difficulties on the manual retrieval and analysis of video information. Computer vision technology has been applied to large-scale video surveillance systems, and video image analysis technology is becoming more automated and intelligent to effectively process massive image information [14].

Video object detection, tracking, and cross-camera recognition are required to execute MCVO extraction. Object detection is conducted to determine the object's location area from the image. The current video object detection methods are mainly based on deep learning, which is divided into two categories: The two-stage model based on region proposal and the one-stage model from end to end. The two-stage model is developed from region-convolutional neural network (R-CNN) and mainly includes fast R-CNN [15], faster R-CNN [16], and mask R-CNN [7], whereas the one-stage model mainly includes the YOLO [8,17] series of methods. Furthermore, object tracking is conducted to predict the size and location of the video object in subsequent video streams under the selection of a tracking object from a given image. Currently, the most widely used tracking methods are filtering [18] and deep learning-based methods [19]. Cross-camera recognition is necessary to determine the spatial–temporal correspondence of video objects from multiple cameras. The implementation methods include the metric-based learning recognition of image features [20] and the local feature-based recognition [21] developed from prior studies. Meanwhile, video sequence [22] and GAN-based recognition methods [23] are developed to introduce learning contents to overcome the insufficiency of the training samples.

In the area of V-GIS integrated data organization, early researchers constructed prototype systems such as multimedia GIS [24], geo-video [25], and video GIS [26] by describing the correspondence between video frames and geographic locations. These studies implemented geographic video image retrieval. In recent years, researchers have focused on the fusion of a projected video image and geographic scene. The geographic video-scene data fusion organization method based on camera spatialization model is formed by constructing image-geographic spatial mapping [2]. Typical camera spatialization models include quadrilateral models applied to 2D planes [27], quadrilateral pyramid models in 3D scenes [28], and grid camera-based coverage analysis models [29]. Researchers developed a series of methods for organizing V-GIS data fusion by geolocation and annotation of video data using the aforementioned models. In some of these methods (e.g., view-based R-tree [3] and camera-based topology indexing [30]), video data organization is analyzed by examining the camera field of view. The other methods used moving object texture association [31], spatial–temporal behavior association [32], and semantic association [33].

A suitable mapping method must be selected to project the video on to the virtual scene model to integrate videos with geospatial information [34,35]. Katkere A. [36] first proposed the concept of fusion expression between video and GIS by using different mapping methods for various semantic-type regions, such as moving objects and video scenes, and building an immersing system based on multi-camera video data. According to different mapping methods, the information fusion methods of surveillance video and virtual scene are divided into two categories: GIS-video image fusion (image projection) [37] and GIS-video moving object fusion (object projection) [38]. The implementation forms of GIS-video image fusion, including video image linked search analysis [4] and videos that are projected to the geographic scene [33], are easy to implement but lack the ability to analyze and understand video image contents. The object projection method extracts video semantic objects from the original video through object detection. Specific implementation methods are divided into three types: (1) foreground and background independent projection [28,39]; (2) foreground projection [31,40]; and (3) foreground abstraction [41]. The foreground and background independent projections project the foreground moving object subgraphs to their corresponding spatial–temporal position in the virtual scene. The foreground projection shows the foreground moving object's subgraph in the corresponding spatial–temporal position in the scene while omitting the projection of the video background image. Foreground abstraction projects the moving object's avatar to the corresponding spatial–temporal position in the geographic scene and simplifies the representation of the video moving object.

3. Fusion between Multiple Camera Objects and GIS

In this section, we first introduce the extraction and data organization for MCVO and then analyze the characteristics of different V-GIS fusion methods and determine the implementation form of the integration of MCVO and GIS.

3.1. Extraction and Data Organization of Multi-Camera Video Moving Objects

Video moving object extraction includes three steps: video moving object detection, tracking, and cross-camera recognition. Figure 2 shows their process relationships. After completing the processes, the original video images are transferred to the data of each video moving objects:

$$C = \left\{ I_j, P_j, (j = 1, 2, \ldots, n) \right\}, \tag{1}$$

where C represents the dates of a video moving object; n is the number of the video frames in which the object occurred; and I_j and P_j represent the location and sub-graph image of the current object in each frame, respectively.

This study adopts the spatial–temporal pipeline (STP) as the basic unit to uniformly describe the MCVO information.

As Obj is the total set of all the moving objects from multiple cameras. $Cube_k$ represents each STP in the current k^{th} camera video range. A total of N_k video moving objects exist in $Cube_k$, and $C_{k,i}$ is the STP of each moving object. $Cube_k$ and $C_{k,i}$ can be expressed as:

$$Obj = \{ Cube_k, (k = 1, 2 \ldots L) \}, \tag{2}$$

$$Cube_k = \left\{ C_{k,i}, (i = 1, 2, \ldots, N_k) \right\}, \tag{3}$$

$$C_{k,i} = \left\{ I_{k,i,j}, P_{k,i,j}, (j = 1, 2, \ldots, n) \right\}, \tag{4}$$

where $I_{k,i,j}$ and $P_{k,i,j}$ represent the geolocation of the i^{th} video moving object in the j^{th} video frame of the camera. This study proposes a data-based MCVO ontology so that the organization of the video moving object data can effectively reflect the association among cross-camera video moving objects, because a single video moving object may appear in many different cameras. This organization realizes the unified organization record of a cross-camera video moving object with the same entity

that appears in different camera videos and requires the merging of the original video moving object's STP. Let L_o denote the total number of video moving objects after merging on cross-camera video moving objects with the same entity, and the expression of the STP of each video moving object is:

$$Obj = \{Cube_i(i = 1, 2, \ldots, L_o)\}(L_o \leq L), \tag{5}$$

$$Cube_i = \left\{C_{k1,i}, C_{k2,i} \ldots C_{ko,i} \ldots, (k1, k2, \ldots ko) \in (1, 2 \ldots L)\right\}, \tag{6}$$

where Obj is the set of all video moving objects in multiple cameras, $Cube_i$ indicates the global STP of the video moving object with serial number i in the surveillance video network, and $C_{ko,i}$ denotes the STP subsequence of the video moving object number i in the ko^{th} camera.

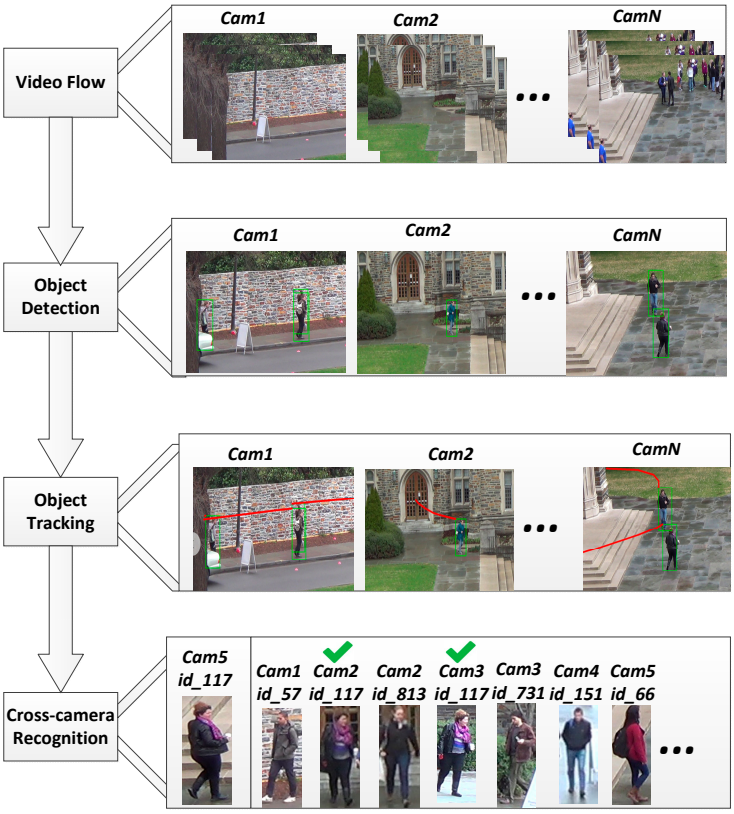


Figure 2. Steps of video moving object extraction.

3.2. Fusion between Video Moving Object and GIS

On the perspective of spatial positioning, the video's field of view was georeferenced to determine the geospatial range of the video image (Figure 3a). The video moving object had to locate the geospatial trajectory to determine the position of the object's subgraph in each video frame (Figure 3b). Both of the fusions were carried out by determining the pairs of points with the same name between the image's geospatial coordinates, solving the camera parameters, and constructing the video's geospatial mapping equation.

The equations for geospatial video mapping were established using the homography matrix method. Figure 4 shows the relationship between the geospatial and the image space coordinate systems. The center of the camera is denoted by C; the image space coordinate system is denoted by $O_i\ X_i\ Y_i$; and the geospatial coordinate system is denoted by $O_g\ X_g\ Y_g\ Z_g$.

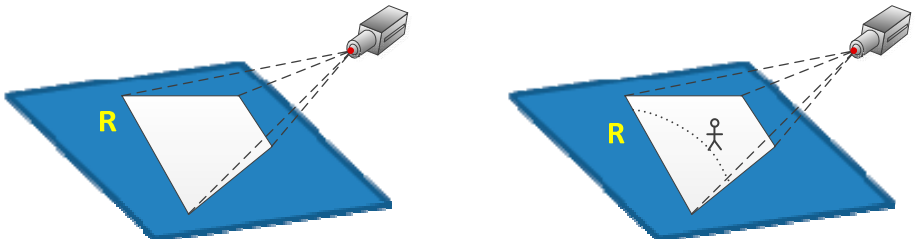


Figure 3. Schematic of spatial mapping. (**a**) The sight region of the camera; (**b**) The position of each moving object in the geographic space.

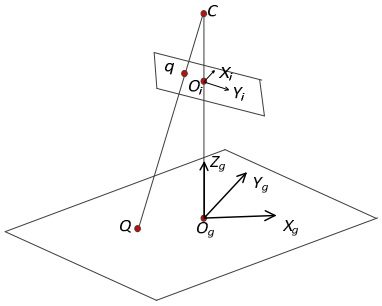


Figure 4. Camera and geospatial coordinate system, image space coordinate system.

Assuming that q and Q are respective points in the image's spatial and geographic coordinate system with similar names, then let the homography matrix be M such that the relationship between q and Q is:

$$q = MQ, \tag{7}$$

where M is represented as follows:

$$M = \begin{bmatrix} k_1 & k_2 & t_x \\ k_3 & k_4 & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{8}$$

M has six unknowns. Thus, at least three pairs of images and geospatial points should be determined to solve M. When M is determined, the coordinates of any point in the geographic space can be solved as:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = M^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{9}$$

The fusion of video moving object and GIS includes three implementations: foreground and background independent projection, foreground projection, and foreground abstraction. We qualitatively analyzed the information expression ability of various video moving object–GIS fusion methods to select the appropriate fusion form of MCVO and GIS. Table 1 shows the results.

Table 1 shows that the foreground and background independent projections simultaneously mapped the background and foreground information of the video, thereby making the video's foreground object less prominent. Meanwhile, foreground abstraction completely abandoned the expression of the original video image data, and no image- and scene-associated expressions were observed. As a result, neither of the two implementations were suitable for the integration of MCVO and GIS. By contrast, foreground projection can express image and scene correlation and represent images spatially. The remaining discussion of this paper on the integration of video moving object and GIS was based on foreground projection because it was suitable for MCVO and geographic scene fusion expression. However, although this problem can be alleviated by optimizing the selection of the

viewpoints to some extent, one disadvantage of foreground projection is that it can only support a certain range of viewpoints displayed in the virtual geographic scene.

Table 1. Analysis of visual feature characteristics of virtual and real expression patterns.

Title 1	Displaying Environment	Ability on Supporting Virtual View Browsing	Correlation Representation Ability between Image and Virtual Scene	Ability on Highlighting Video Foreground Object
Image projection	2D/3D	Range view	Yes	No
Foreground and background independent projection	3D	Range view	Yes	No
Foreground projection	3D	Range view	Yes	Yes
Foreground Abstraction	2D/3D	Arbitrary view	No	Yes

3.3. Data Organization of Spatial–Temporal Trajectory

A single camera can only record the spatial–temporal trajectories of the moving objects within the field of view of this camera. While the fields of view of different cameras are not continuous, for analyzing the trajectory of a video moving object in a multi-camera environment, it is necessary to re-organize the trajectory of the same object occurred in different cameras.

The processing steps came as follows: using the automatic object detection algorithm developed from computer vision to detect the video moving objects, extract the positions and sub-pictures, then use the tracking algorithm to generate the spatial–temporal trajectories of moving objects. Based on the mapping model described in Section 3.2, we can get the instantaneous position of the moving object in each frame. In each camera's field of view, all instantaneous spatial positions of a moving object are temporal aligned, combining into the local trajectory of this object in the current camera. Since the field of view of different cameras in geospatial space is discontinuous, it is necessary to use the moving object re-recognition algorithm to perform cross-camera recognition to obtain the global trajectory of each video moving object in a multi-camera environment (as shown in Figure 5).

In order to effectively reorganize the association of different levels of video moving object trajectory, Obj is the total set of all moving objects in the geographical scene, and there are N_k -th moving objects in the k-th camera field of view, and the local trajectory of each dynamic object within the camera's field of view is $C_{k,i}$. The expression of Obj and $C_{k,i}$ are as follows:

$$Obj = \left\{C_{k,i}, (k = 1, 2 \ldots L)(i = 1, 2, \ldots, N_k)\right\} \tag{10}$$

$$C_{k,i} = \left\{P_{k,i,j}, (j = 1, 2, \ldots, n)\right\}, \tag{11}$$

where $I_{k,i,j}$ and $P_{k,i,j}$ respectively represents the geospatial position of the i-th moving object in the j-th video frame in the k-th camera. Since the same moving object may appear in different camera fields, in order to express the moving object cross-camera association, L_o represents the actual total number of moving objects in the geographic scene, and $Cube_i$ is the global trajectory of each moving object. The expression of $Cube_i$ and Obj are as follows:

$$Cube_i = \left\{C_{k1,i}, C_{k2,i} \ldots C_{ko,i} \ldots, (k1, k2, \ldots ko) \in (1, 2 \ldots L)\right\}, \tag{12}$$

$$Obj = \{Cube_i (i = 1, 2, \ldots, L_o)\}(L_o \leq L), \tag{13}$$

where $Cube_i$ represents the global trajectory of the i-th moving object in the geographical scene, and the local trajectory of the moving object in the $k1, k2, \ldots ko$ -th camera is $C_{k1,i}, C_{k2,i} \ldots C_{ko,i}$. Obj is still the total set of all moving objects in the geographical scene.

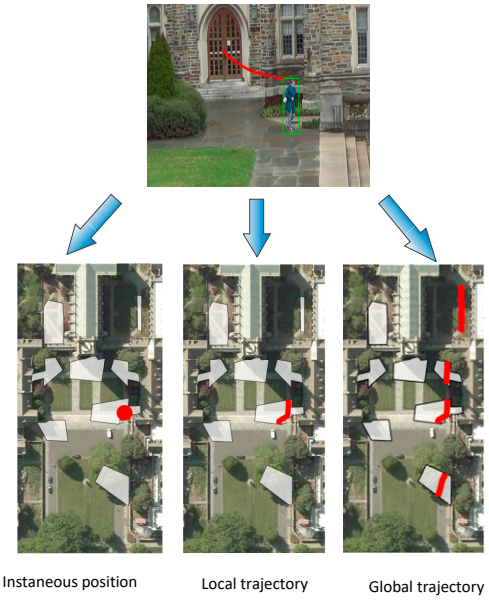


Figure 5. Moving object instantaneous position, local trajectory, global trajectory.

4. Architecture of GIS-MCVO Surveillance System

The video surveillance system was integrated with GIS, and the prototype of GIS-MCVO system was designed and developed on the basis of data organization and MCVO spatialization. The system assists the users to achieve rapid reorganization and effective understanding of multi-camera video content and geospatial association in the urban environment, which are evacuated by storing geospatial information, surveillance video, and video moving object information separately and performing fusion display and comprehensive analysis.

4.1. Design Schematic of the System

The overall system design of GIS-MCVO followed the framework of service-oriented software architecture. The framework of the system was divided into a function layer, data layer, service layer, business layer and representation layer from bottom to up, as shown in Figure 6.

(1) **Function layer:** The function layer is a server with data processing and analysis functions. This layer is used for pre-processing GIS and video data and comprises functional modules for video data acquisition, video moving object extraction, video data geospatial mapping, and cross-camera object recognition. In addition, the function layer can provide basic data support for real-time publishing.

(2) **Data layer:** This layer is supported by the database and is mainly used to store, access, and manage geospatial, video image, and video moving object data and to provide data services to clients.

(3) **Service layer:** The service layer publishes the data service of the underlying system database, including video stream image, video moving object, and geospatial information data services. This layer provides real-time multisource data services to terminal users and remote command centers.

(4) **Business layer:** The business layer selects relevant data service content according to the demand of the system user. Through analysis, this layer fetches different services and generates and transmits the corresponding result to the representation layer.

(5) **Representation layer:** In the representation layer, the user can apply multiple modes on the MCVO and GIS fusion, along with related functions on application and analysis by using a common browser under various operating system platforms.

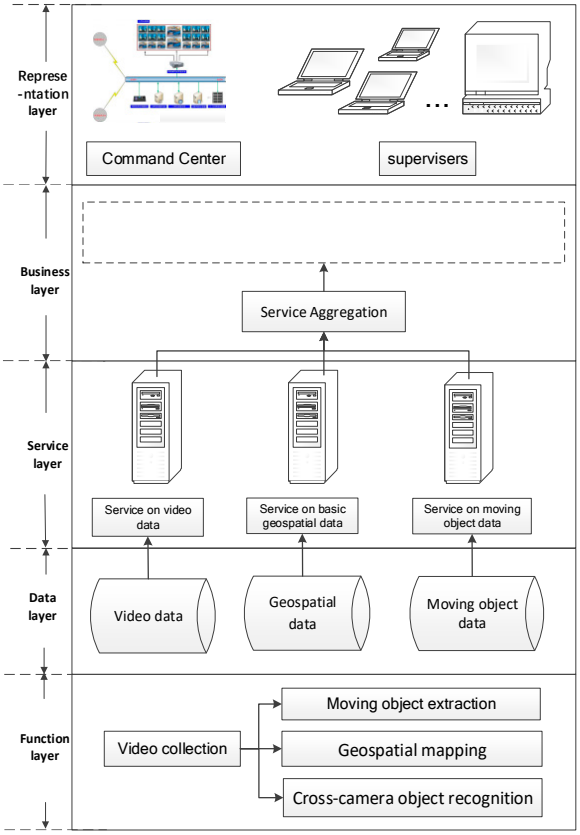


Figure 6. Design Schematic of the system.

4.2. Design of System Functions

This section describes the modules in the function layer and their functional support relationships (Figure 7).

(1) **Moving object extraction module:** This module uses detection and tracking algorithms to extract moving objects; separate the video's foreground and background; achieve cross-camera recognition on objects from different cameras; and stores the trajectory, type, set of sub-graphs, and other associated information of the moving objects.

(2) **Video spatialization module:** This module constructs the mapping matrix by selecting the associated image and geospatial mapping model and calibrates the internal and external parameters of the camera for video spatialization.

(3) **Virtual scene generation module:** This module is mainly used to load the virtual geographic scene, virtual point of view, position of the surveillance camera, and sight of video image. This module builds the foundation of fusion representation. Many applications based on GIS-MCVO system are applied under the condition of establishing this module.

(4) **Moving object spatial–temporal analysis module:** To achieve some specific applications, this module synthesizes the related information on the video moving objects and the geographic scene. This module also obtains the necessary result to be outputted in the representation module.

(5) **Fusion representation module:** This module selects the fusion pattern between the moving objects and the virtual geographic scene by performing visual loading on video images, moving object trajectory, sub-graphs, and spatial–temporal analysis results.

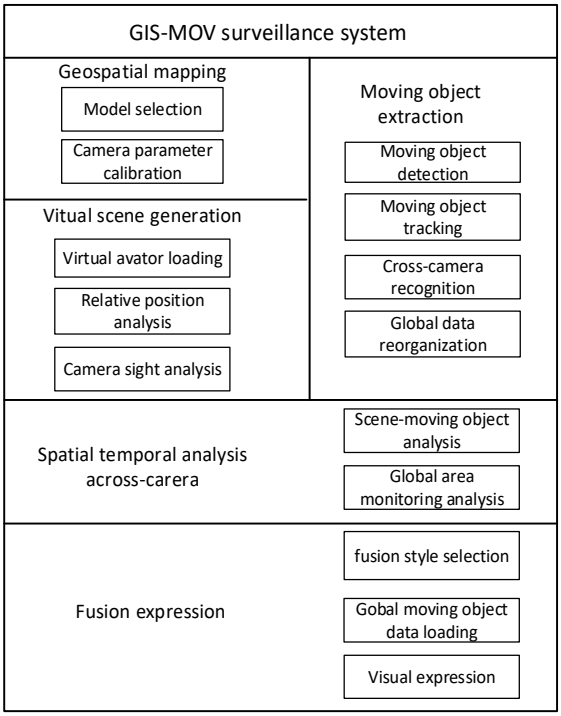


Figure 7. Diagram of system functions.

5. Applications and Potential Benefits for GIS-MCVO Surveillance System

In this chapter, we have briefly described the applications and potential benefits of the GIS-MCVO surveillance system and compared the results with the evacuation results of traditional and GIS-video image surveillance systems. The implementation of GIS-MCVO system is shown in Figure 8.

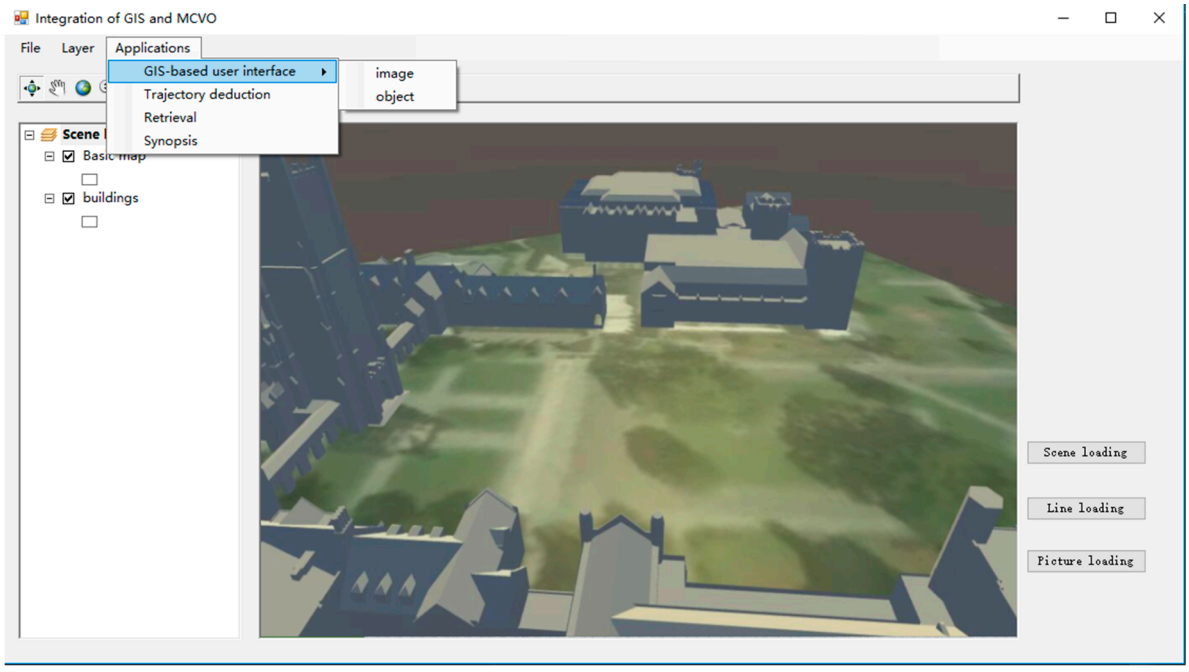
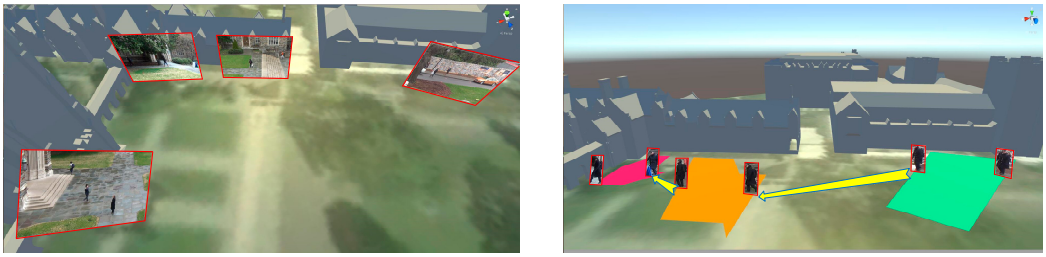


Figure 8. Implementation of GIS-MCVO system.

5.1. GIS-Based User Interface

The user interface of a traditional video surveillance system includes a monitor that displays video from a selected camera, which can switch among different cameras or simultaneously display

multiple channels of video in a grid interface. As the number of cameras increases, the visual interface architecture becomes less usable. When users need to retrieve and comprehensively analyze multi-camera video information with complex spatial relationships, the situation becomes unfavorable because the grid interface cannot express the spatial relationship of the cameras nor can it express video object information in different cameras. Effective identification requires three factors: familiarity of the users on the camera's shooting direction; spatial relationship to the camera's field of view; and the ability to manually and rapidly complete camera selection, switching operations, and moving object tracking tasks. These operations require long-term training and experience, and even if the user can effectively master them, mistakes are still inevitable. The defects of the grid interface of multi-camera videos include: (1) the lack of ability to express a spatial relationship among camera views; and (2) the lack of ability to extract video moving object information and cross-camera identification. The video image and GIS integrated visualization system, can effectively solve the defect described in (1), but did not solve the defect stated in (2) because the system still needs to manually search and analyze moving objects in long-term and large-data scaled video and artificially determine the spatial relationship of the video objects from different cameras due to the lack of extraction and analysis processes on video moving objects. Therefore, the GIS and video image integrated visualization system still experienced considerable manual processing pressure when dealing with actual multi-camera video monitoring tasks as object retrieval and behavior analysis (Figure 9a).



(**a**) (**b**)

Figure 9. Fusion diagram between GIS and multi-camera videos. (**a**) GIS and video image integrated visualization. (**b**) GIS and MCVO integrated visualization.

To reduce the user's information retrieval pressure on surveillance video and promote the expression effectiveness of a video moving object and 3D geographic scene model, the GIS-MCVO system, in which the geospatial visualization of video information was achieved by dynamically expressing the video object subgraph in a geographic scene, and was constructed on the basis of video object trajectory and sub-graph data. The advantages of this visual interface were presented as follows: (1) Only the video moving object's sub-graph is expressed, thereby greatly reducing the amount of video information that the user needs to retrieve and watch; (2) The video object sub-graph is displayed in a 3D geographic scene using a planar map, thereby avoiding the video image texture distortion–alignment problem in 3D scene model fitting, reducing the amount of image rendering calculation, and improving the efficiency of video information expression; and (3) the GIS-MCVO system can synchronously and relatedly express the MCVO trajectory by adding the view polygon, identifying the same entity object, and optimizing a virtual viewpoint because the same entity video object trajectory is associated with multiple spatial–temporal cameras (Figure 9b).

In summary, the proposed GIS-MCVO system can demonstrate the spatialization of video image and video object trajectory. In addition, the system can not only effectively express the spatial relationship within the camera's field of view but can also bring about the visual representation of the video information based on the video object trajectory and sub-graph data, thereby resolving the defects of the grid interface of multi-camera videos. The GIS-MCVO system eliminates the manual processing of multi-camera surveillance video information.

5.2. Video Compression Storage

The GIS-MCVO system only stores the sub-graph and spatial–temporal trajectory information of the video moving object, but not the background information. Therefore, when storing and analyzing video moving objects, video data compression, also known as VR fusion video compression, occurs [42]. This method converts video data from an image level to an object level. The obtained compressed data constitutes a series of correlated compression levels in different combinations (Figure 10).

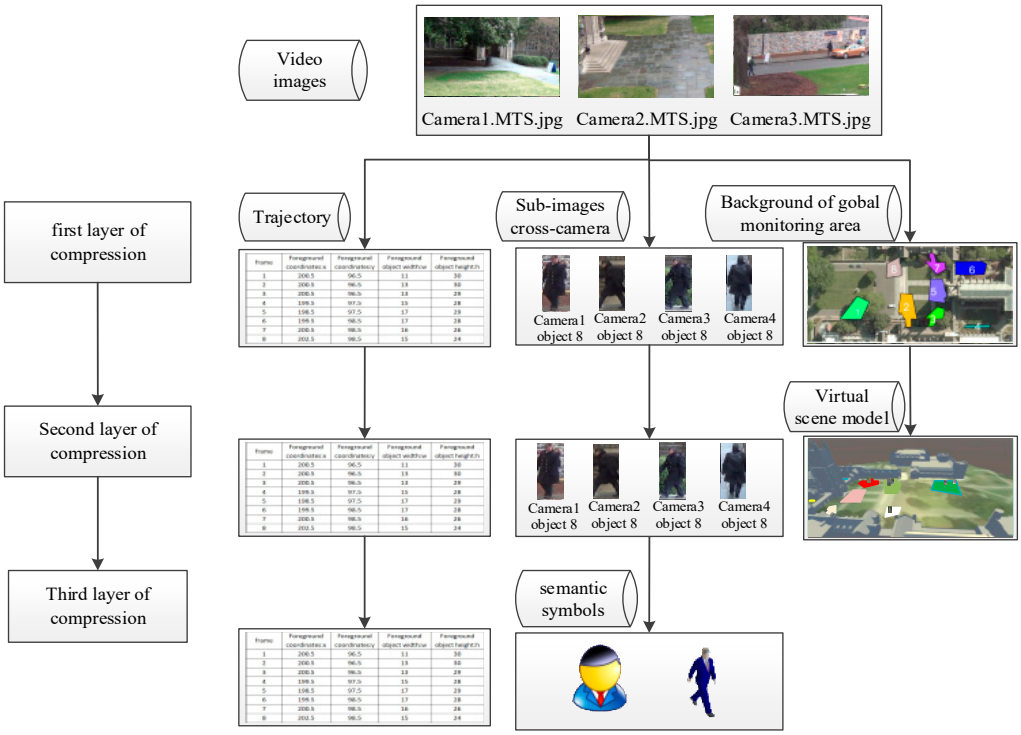


Figure 10. Video compression data hierarchical relationship diagram.

From the perspective of a compression mechanism, video image compression is achieved by constructing predictive models in accordance with the H.264 standard, which predicts video image pixels via intra- or inter-frame prediction. Video image compression aims to reconstruct the original video by using compressed data. Thus, the capability of recovering the original video images should be considered. Furthermore, VR fusion video compression represents video information in simplified approaches (e.g., showing only the sub-graphs or avatars of the moving object). As a result, the ability of the latter to recover the original video image does not need to be considered. In terms of the data compression effect, the video data used in the object projection fusion patterns have data compression relations with the original video sequence images. Furthermore, data compression relations exist among the three patterns of object projection fusion.

First layer of compression: In this layer, the compressed data were oriented to the foreground and background independent projection patterns. The sub-graphs of the moving objects, spatial–temporal position, and background images were extracted and stored separately. This compression layer converted video information from the image level to the object level.

Second layer of compression: In this layer, the compressed data were oriented to the foreground projection pattern, and the virtual scene model was used instead of the video background. This compression layer transferred the background that represented the camera view from the image to the virtual scene model.

Third layer of compression: In this layer, the compressed data were oriented to the abstract of the foreground projection pattern. The virtual avatar in a semantic symbol was used instead of the

sub-graphs to display video moving objects in a virtual geographic scene, and the spatial–temporal position was the only information that needed to be stored.

To test the compression efficiency of the data for storage, we examined a set of video images and recorded the trend of the compression rate K_i with respect to the number of input video frames for the different layers. The experimental results are presented as follows:

In Figure 11, the magnitudes of compression rate in the first layers as K_1, was in the order of 10^{-2},whereas the magnitude of compression in the second and third layer as K_2 and K_3, were in the order of 10^{-4}. These results proved that the video compression process in the GIS-MCVO system could effectively reduce the amount of video data.

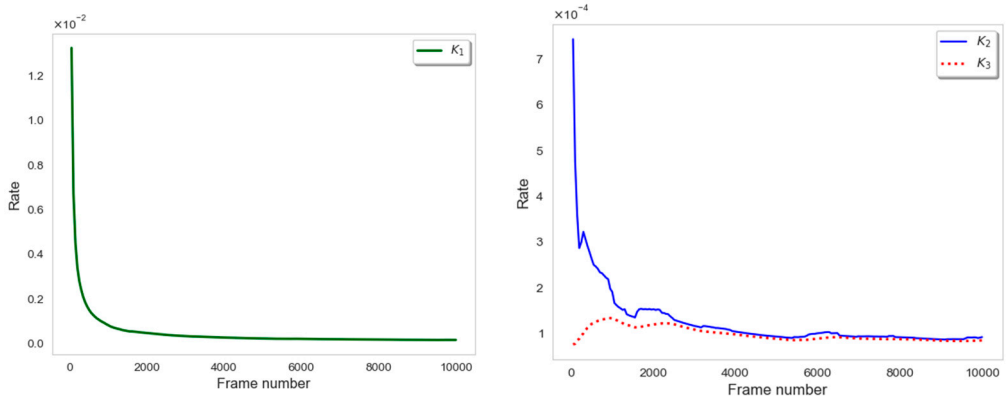


Figure 11. Value of K_i changed with frame number. (**a**) K_1; (**b**) K_2 = 2 (solid line) and K_3 = 3 (dotted line).

5.3. Trajectory Deduction in Visual Blind Zone

In the GIS-MCVO system, the trajectory of an arbitrary video object, which appears in a plurality of the camera's fields of view, can be determined through cross-camera video object recognition. However, factors, such as the size of the camera field of view, the number of cameras, and the size of the monitoring area limit the trajectory. Several visual blind zones, in which visual information cannot be captured by any camera, exist, but the object's general motion trends and paths in these zones can be determined on the basis of camera spatial relationships, road network, and 3D geographic scene model (Figure 12). Through trajectory deduction in visual blind zones, the global trajectory in multi-camera surveillance areas can be estimated and supplemented by the implementation of specific functions in GIS-MCVO as video object retrieval, video synopsis, and video object behavior understanding.

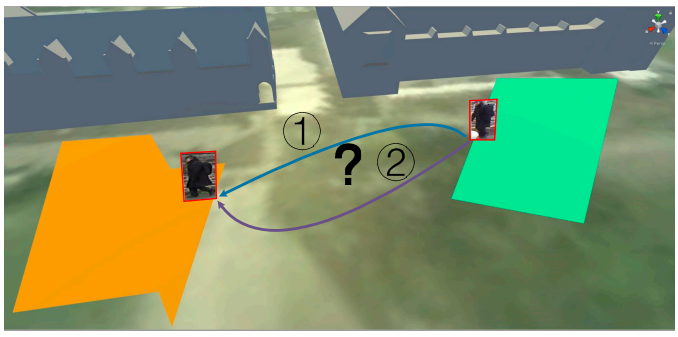


Figure 12. Trajectory deduction in the camera's visual blind zone.

5.4. Retrieval of MCVO

Even when a user detects a large amount of video moving objects, they will always be interested in several video moving objects with specific conditions. Thus, video moving object retrieval is needed.

The search conditions are divided into two types: appearance and trajectory descriptions. In this section, we discuss the problems of the trajectory description of MCVO retrieval.

The input modes of trajectory description-based video object retrieval include moving object instance, graphical description, and semantical description. The trajectory graphical description is divided into area of interest (AOI)-based type and trajectory template-based type. The topological relationship between the search object and the AOI or template trajectory is used to determine whether the search condition is met. The research on traditional trajectory description is limited by generating AOI or trajectory template in the image space [43,44]. Moreover, realizing geospatial comprehensive retrieval on video object trajectory is impossible in a multi-camera environment. Thus, the search work was divided and refined, and the retrieval analysis was processed individually in each camera to obtain the search results. In recent years, some studies have attempted to construct an AOI [45] and trajectory template [46] in geographic space by combining multi-camera fields of view to carry out the spatialization of trajectory retrieval condition description. Although these methods realize the geospatial analysis of the retrieval conditions, the video content of each camera must be retrieved separately due to the lack of spatial–temporal correlation analysis of different camera video object trajectories. To solve the problems mentioned above, the proposed GIS-MCVO system had the function of integrated representation and analysis of geospatial video moving object. The video object retrieval function in GIS-MCVO can analyse the video object trajectory retrieval condition globally, thereby deviating from the limitations of the camera lens. Video object sub-graphs can be integrated in the 3D scene model to express the retrieval results (Figure 13). The camera-by-camera search and the global search can be realized by sketching the trajectory template in the virtual geographic scene of the GIS-MCVO system. The camera-by-camera search analyzes the spatial relationship between the search condition and the camera's field of view, and then matches the search condition with the trajectory of video moving object in each camera's field of view and returns the results of the query. The global search directly matches the search condition with the cross-camera global trajectory of video moving objects and returns the results of the query. The search results are as shown in Figure 14.

Figure 13. Video object retrieval in geospatial multi-camera.

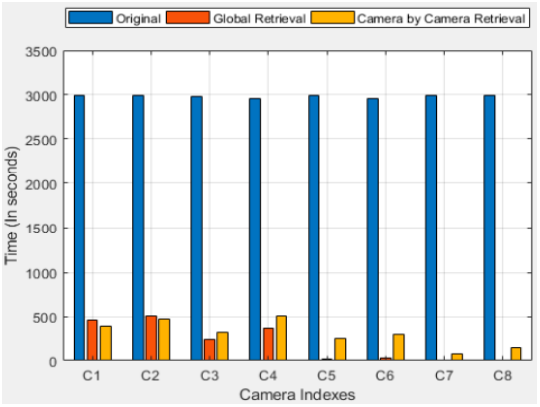


Figure 14. Time to watch videos per camera.

5.5. Synopsis of Multiple Videos

The video moving objects are timely sparse distributed in the original video. If the video moving objects are watched after the original time, then a blank screen, which is not conducive to the efficient expression of the video and wastes the user's working time, might be displayed for a long duration. In response to this problem, several research studies have attempted to increase the temporal displaying density of the video moving object by reducing the video playback duration; this method is referred to as video synopsis [47]. Video synopsis changes the playback sequence of the video moving object according to their spatial–temporal relationships and concentrates on expressing large amounts of video moving objects in a short period of time while these video moving objects appear in different time segments in the original video (Figure 15). The image platform is utilized to generate video synopsis by creating a short video. Although this generation mode is applied in multi-camera video synopsis, it produces problems in two aspects: First, the optimization in image generation appears as an overlap of the video moving object sub-graphs and as continuous updates of the video background image; and second, the spatial information expression among different videos appears as an inability to express spatial–temporal associations among the video objects from different cameras and optimized selection in virtual view. At present, existing research focuses on fusion processing of video information in different cameras using image matching [48], camera angle optimization [49], and video image cropping fusion. These studies only consider the correlation of pixel content among video images but fail to effectively solve the spatial information expression problems, such as MCVO correlation expression and optimized selection in virtual view.

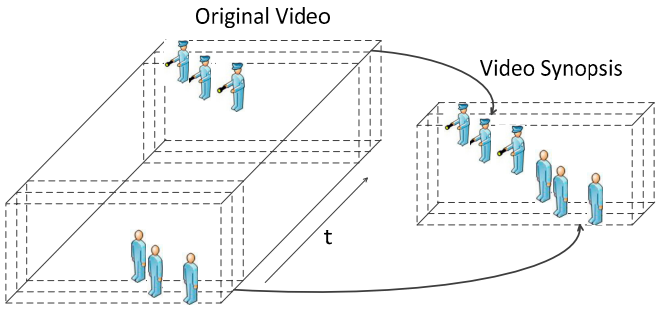


Figure 15. Schematic diagram of video synopsis.

To overcome the problems above, the proposed GIS-MCVO system was designed on the basis of video moving object extraction and cross-camera correlation to flexibly select the video moving objects and cameras, thereby effectively expressing the spatial–temporal information of a large amount of MCVO in the geographic scene and achieving multi-camera video synopsis. GIS-MCVO-based multi-camera video synopsis concentration converts from the object level, in which the video moving

object's playback time is adjusted and is extended to the camera level. This level achieves integrated operations, including camera lens switching and cross-camera video moving object playback time reset (Figure 16).

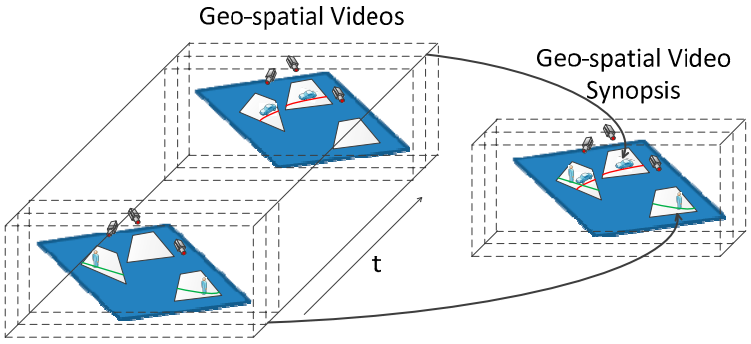
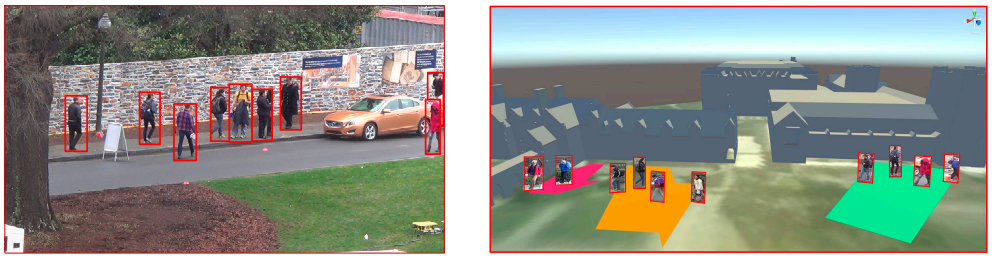


Figure 16. Schematic diagram of multi-camera Geo-spatial video synopsis.

The proposed GIS-MCVO system expresses the video object as a 3D geographic scene model. This mode not only avoids image optimization problems, such as overlapping of video objects and continuous updating of the video background however, can also optimize the selections on object display and camera display sequences along with its effective expression on the spatial–temporal association of video objects among different cameras (Figure 17).



(**a**) (**b**)

Figure 17. Comparison of effects between different kinds of video synopsis. (**a**) Single camera Video synopsis in image. (**b**) Multi-camera geo-spatial video synopsis.

In order to analyze the duration reduction effect of video synopsis, the following function was used to calculate the compression ratio of video frames of multiple cameras.

$$\eta = [(n_i)t_0 + max(f_w)]/F, \tag{14}$$

where n_i represents the number of video moving objects; t_0 represents starting interval frames between each two adjacent video moving objects; $\max(f)$ represents the largest number of frames among all the video moving objects; F represents the summary of the original video frames from all the cameras. Under different values of t_0, the compression ratio of video frames is shown in Figure 18.

As can be seen in Figure 18, video synopsis based on GIS-MCVO can reduce the video play-back time; on the other hand, the compression ratio obtained after video object cross-camera recognition is lower than that without cross-camera recognition, which proves that object cross-camera recognition can improve the efficiency of the expression of MCVO.

In summary, GIS-MCVO-based video synopsis can assist users in rapidly retrieving the spatial–temporal trajectory and image information of many video moving objects in a multi-camera environment, along with efficiently understanding the spatial–temporal behavior of MCVO.

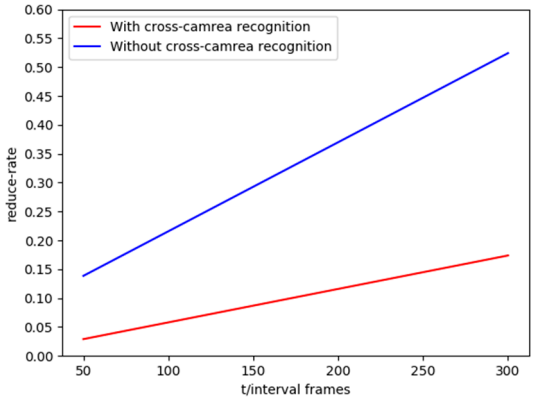


Figure 18. Frame compression ratio in multi-camera synopsis.

6. Conclusions

This work mainly discussed the integration of MCVO and GIS. The implementation of the integration was carried out on the basis of V-GIS fusion [9], V-GIS practical application [1,3], and our preliminary work on single camera video object and GIS integration [10]. The advantage of this integration was that it could achieve not only the extraction of video key information and spatial correlation visualization but also the spatially associated analysis of multi-camera video object, thereby assisting users to effectively monitor video operations. The main contributions of this paper are presented as follows:

(1) The research on single-camera video object and GIS integration is extended to MCVO and GIS integration along with the consideration of the spatial–temporal correlation among different camera objects. In addition, the conceptual fusion model between MCVO and GIS is proposed.

(2) The GIS-MCVO system is built on the basis of constructing multi-camera video moving objects and GIS integration related data organization. The overall architecture and function design of the system are presented and analyzed.

(3) This paper analyzed the related applications of the GIS-MCVO system by comparing the traditional image-based surveillance system with GIS-video image integrated surveillance system. The results showed that GIS-MCVO is advantageous in the applications of system user interface, video compression storage, trajectory deduction, video object retrieval, and video synopsis.

GIS-MCVO realizes data organization, spatial–temporal analysis, and visual expression of MCVO integrated with GIS. Compared with GIS-video image fusion, the fusion between GIS and MCVO has the following advantages: (1) multi-camera video object and the geographic scene are integrated for expression and analysis, rather than video image–GIS integrated fusion; (2) video moving object from different cameras with the same entity are integrated, analyzed and expressed in GIS; and (3) the GIS-based user interface assists users in retrieving and analyzing video moving objects quicker and more concisely; and (4) the compression rate of video data is in the order of 10^{-2} to 10^{-4} by achieving the integration. However, the fusion between GIS and MCVO has the following disadvantages: (1) Loss of video information, in which some visual image information of video objects are abandoned during evacuation; and (2) uncertainty on video object due to the missed or false detection of the video object. Moreover, the visual blind zone between the cameras, which decreases the accuracy of the extracted MCVO results in multi-camera video moving objects' trajectories, that are not necessarily accurate.

We implemented a prototype of the GIS-MCVO surveillance system, which can be used to illustrate the proof of integration of GIS-video object integrated analysis or considered a platform to carry out the interactive monitoring and analysis of video and geographic information by adding other functions in the future. In further research, we will introduce GPS real-time positioning data, remote sensing images, and lyric data in the integration of video and GIS by investigating the geographic video understanding and analysis supported by multisource information.

Author Contributions: Conceptualization, Yujia Xie; methodology, Yujia Xie, MeiZhen Wang; software, Bo Mao; validation, Feiyue Wang; formal analysis, MeiZhen Wang; investigation, Yujia Xie, Xuejun Liu; resources, Xuejun Liu; data curation, Bo Mao; writing—original draft preparation, Yujia Xie; writing—review and editing, Yujia Xie, Feiyue Wang; visualization, Feiyue Wang; supervision, MeiZhen Wang, Xuejun Liu; project administration, MeiZhen Wang; funding acquisition, Yujia Xie, MeiZhen Wang, Feiyue Wang.

Funding: This research was funded by National Natural Science Foundation of China (NSFC) under Grant 41671457, 41771420, and 41801305.

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

References

1. Milosavljević, A.; Rančić, D.; Dimitrijević, A.; Predić, B.; Mihajlović, V. Integration of GIS and video surveillance. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 2089–2107. [CrossRef]
2. 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]
3. Wu, C.; Zhu, Q.; Zhang, Y.T.; Du, Z.Q.; Zhou, Y.; Xie, X.; He, F. An Adaptive Organization Method of Geovideo Data for Spatio-Temporal Association Analysis. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *2*, 29. [CrossRef]
4. Han, Z.; Cui, C.; Kong, Y.; Qin, F.; Fu, P. Video data model and retrieval service framework using geographic information. *Trans. GIS* **2016**, *20*, 701–717. [CrossRef]
5. Xie, X.; Zhu, Q.; Zhang, Y.; Zhou, Y.; Xu, W.; Wu, C. Hierarchical semantic model of Geovideo. *Acta Geod. Cartogr. Sin.* **2015**, *44*, 555–562.
6. Kong, Y. Design of Geo Video Data Model and Implementation of Web-Based VideoGIS. *Geomat. Inf. Sci. Wuhan Univ.* **2010**, *35*, 133–137.
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
8. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Milosavljević, A.; Dimitrijević, A.; Rančić, D. GIS-augmented video surveillance. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 1415–1433. [CrossRef]
10. 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]
11. 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.
12. Tilley, N. *Understanding Car Parks, Crime, and CCTV: Evaluation Lessons from Safer Cities*; Home Office Police Department: London, UK, 1993.
13. Brown, B. *CCTV in Town Centres: Three Case Studies*; Home Office, Police Department: London, UK, 1995.
14. Huang, K.Q.; Chen, X.T.; Kang, Y.F.; Tan, T.N. Intelligent visual surveillance: A review. *Chin. J. Comput.* **2015**, *38*, 1093–1118.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
18. Lukezic, A.; Vojir, T.; Cehovin Zajc, L.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6309–6318.
19. Valmadre, J.; Ertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H. End-to-end representation learning for correlation filter based tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2805–2813.

20. Liu, H.; Feng, J.; Qi, M.; Jiang, J.; Yan, S. End-to-end comparative attention networks for person re-identification. *IEEE Trans. Image Process.* **2017**, *26*, 3492–3506. [CrossRef] [PubMed]
21. Wei, L.; Zhang, S.; Yao, H.; Gao, W.; Tian, Q. Glad: Global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 420–428.
22. Liu, H.; Jie, Z.; Jayashree, K.; Qi, M.; Jiang, J.; Yan, S.; Feng, J. Video-based person re-identification with accumulative motion context. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2788–2802. [CrossRef]
23. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
24. Charou, E.; Kabassi, K.; Martinis, A.; Stefouli, M. Integrating multimedia GIS technologies in a recommendation system for geotourism. In *Multimedia Services in Intelligent Environments*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 63–74.
25. McDermid, G.J.; Franklin, S.E.; LeDrew, E.F. Remote sensing for large-area habitat mapping. *Prog. Phys. Geogr.* **2005**, *29*, 449–474. [CrossRef]
26. Navarrete, T.; Blat, J. VideoGIS: Segmenting and indexing video based on geographic information. In Proceedings of the 5th AGILE Conference on Geographic Information Science, Palma de Mallorca, Spain, 25–27 April 2002.
27. Walton, S.; Berger, K.; Ebert, D.; Chen, M. Vehicle object retargeting from dynamic traffic videos for real-time visualisation. *Vis. Comput.* **2014**, *30*, 493–505. [CrossRef]
28. Du, R.; Bista, S.; Varshney, A. Video fields: Fusing multiple surveillance videos into a dynamic virtual environment. In Proceedings of the 21st International Conference on Web3D Technology, Anaheim, CA, USA, 22–24 July 2016; ACM: New York, NY, USA, 2016; pp. 165–172.
29. Wang, M.; Liu, X.; Zhang, Y.; Wang, Z. Camera coverage estimation based on multistage grid subdivision. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 110. [CrossRef]
30. Cho, Y.J.; Park, J.H.; Kim, S.A.; Lee, K.; Yoon, K.J. Unified framework for automated person re-identification and camera network topology inference in camera networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2601–2607.
31. Jian, H.; Liao, J.; Fan, X.; Xue, Z. Augmented virtual environment: Fusion of real-time video and 3D models in the digital earth system. *Int. J. Digit. Earth* **2017**, *10*, 1177–1196. [CrossRef]
32. Loy, C.C.; Xiang, T.; Gong, S. Time-delayed correlation analysis for multi-camera activity understanding. *Int. J. Comput. Vis.* **2010**, *90*, 106–129. [CrossRef]
33. Mehboob, F.; Abbas, M.; Rehman, S.; Khan, S.A.; Jiang, R.; Bouridane, A. Glyph-based video visualization on Google Map for surveillance in smart cities. *EURASIP J. Image Video Process.* **2017**, *1*, 28. [CrossRef]
34. De Haan, G.; Scheuer, J.; de Vries, R.; Post, F.H. Egocentric navigation for video surveillance in 3D virtual environments. In Proceedings of the 2009 IEEE Symposium on 3D User Interfaces, Lafayette, LA, USA, 14–15 March 2009; pp. 103–110.
35. Wang, Y.; Bowman, D.A. Effects of navigation design on Contextualized Video Interfaces. In Proceedings of the 2011 IEEE Symposium on 3D User Interfaces (3DUI), Singapore, 19–20 March 2011; pp. 27–34.
36. Katkere, A.; Moezzi, S.; Kuramura, D.Y.; Kelly, P.; Jain, R. Towards video-based immersive environments. *Multimed. Syst.* **1997**, *5*, 69–85. [CrossRef]
37. Lewis, P.; Fotheringham, S.; Winstanley, A. Spatial video and GIS. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 697–716. [CrossRef]
38. Wang, X. Intelligent multi-camera video surveillance: A review. *Pattern Recognit. Lett.* **2013**, *34*, 3–19. [CrossRef]
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 City, QC, Canada, 27–30 September 2015; pp. 2875–2879.
40. Baklouti, M.; Chamfrault, M.; Boufarguine, M.; Guitteny, V. Virtu4D: A dynamic audio-video virtual representation for surveillance systems. In Proceedings of the 2009 3rd International Conference on Signals, Circuits and Systems (SCS), Medenine, Tunisia, 6–8 November 2009; pp. 1–6.
41. Pan, C.; Chen, Y.; Wang, G. Virtual-Real Fusion with Dynamic Scene from Videos. In Proceedings of the 2016 International Conference on Cyberworlds (CW), Chongqing, China, 28–30 September 2016; pp. 65–72.

42. Collins, R.T.; Lipton, A.J.; Fujiyoshi, H.; Kanade, T. Algorithms for cooperative multisensor surveillance. *Proc. IEEE* **2001**, *89*, 1456–1477. [CrossRef]
43. Katz, B.; Lin, J.J.; Stauffer, C.; Grimson, W.E.L. Answering Questions about Moving Objects in Surveillance Videos. In *New Directions in Question Answering*; Springer: Dordrecht, The Netherlands, 2003; pp. 145–152.
44. Hu, W.; Xie, D.; Fu, Z.; Zeng, W.; Maybank, S. Semantic-based surveillance video retrieval. *IEEE Trans. Image Process.* **2007**, *16*, 1168–1181. [CrossRef] [PubMed]
45. Deng, H.; Gunda, K.; Rasheed, Z.; Haering, N. Retrieving large-scale high density video target tracks from spatial database. In Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications, Washington, DC, USA, 1–3 July 2012; p. 19.
46. Panta, F.J.; Qodseya, M.; Péninou, A.; Sèdes, F. Management of Mobile Objects Location for Video Content Filtering. In Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia, Yogyakarta, Indonesia, 19–21 November 2018; pp. 44–52.
47. Pritch, Y.; Rav-Acha, A.; Peleg, S. Nonchronological video synopsis and indexing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1971–1984. [CrossRef]
48. Zhu, X.; Liu, J.; Wang, J.; Lu, H. Key observation selection-based effective video synopsis for camera network. *Mach. Vis. Appl.* **2014**, *25*, 145–157. [CrossRef]
49. Zhang, Z.; Nie, Y.; Sun, H.; Lai, Q.; Li, G. Multi-video object synopsis integrating optimal view switching. In Proceedings of the SIGGRAPH Asia 2017 Technical Briefs, Bangkok, Thailand, 27–30 November 2017; p. 17.