Erschienen in: 9th International Conference on Imaging for Crime Detection and Prevention (ICDP-2019). - Piscataway, NJ : IEEE, 2019. - S. 44-49. - ISBN 978-1-83953-109-5
https://doi.org/10.1049/cp.2019.1166

Computer Vision Meets Visual Analytics: Enabling 4D Crime Scene Investigation from Image and Video Data

Thomas Pollok[⋆], Matthias Kraus[†], Chengchao Qu[⋆], Matthias Miller[†], Tobias Moritz[⋆], Timon Kilian[†], Daniel Keim[†], Wolfgang Jentner[†]

[⋆]Fraunhofer IOSB, Fraunhoferstr. 1, 76131 Karlsruhe, Germany firstname.lastname@iosb.fraunhofer.de
[†]Universität Konstanz, Universitätsstr. 10, 78465 Konstanz, Germany firstname.lastname@uni.kn

Keywords: 4D Reconstruction, Computer Vision, Crime Scene Investigation, Visual Exploration, Forensics

Abstract

In case of a crime or terrorist attack, nowadays much video footage is available from surveillance and mobile cameras recorded by witnesses. While immediate results can be crucial for the prevention of further incidents, the investigation of such events is typically very costly due to the human resources and time that are needed to process the mass data for an investigation. In this paper, we present an approach that creates a 4D reconstruction from mass data, which is a spatio-temporal reconstruction computed from all available images and video footage. The resulting 4D reconstruction gives investigators an intuitive overview of all camera locations and their viewing directions. It provides investigators the ability to view the original video or image footage at any specific point in time. Combined with an innovative 4D interface, our resulting 4D reconstruction enables investigators to view a crime scene in a way that is similar to watching a video where one can freely navigate in space and time. Furthermore, our approach augments the scene with automatic detections and their trajectories and enrich the crime scene with annotations serving as clues.

1 Introduction

Crime scene investigation often remains a lengthy and costly process due to the nature of the massive amount of available data from different sources such as surveillance cameras and mobile footage. However, at the moment, it lacks useful tools for crime scene investigators to efficiently and effectively explore and analyze the mass data, so that it quickly exceeds their capacities sitting in front of video walls to examine through the footage. The main objective of this paper is therefore to create a system that allows for the analysis of a massive number of image and video data with the aid of emerging algorithmic approaches from the field of computer vision combined with intelligent interaction concepts, in order to display all these data in an intuitive and coherent visual 3D representation.

Towards this goal, a number of challenges must be addressed first, including:

- spatio-temporal referencing of massive image and video data captured in and around the crime scene;
- reconstruction of 3D static and dynamic scene parts;
- intuitive representation of the reconstruction results;
- explorative visual analysis;
- convenient user interactions.

In the first stage of the workflow, we conduct spatial referencing of all input data using content-based analysis of the image, and video frames. The 3D reconstruction produces a static scene that is further processed. An optional semi-automated geo-registration of the 3D model onto the world atlas can then be carried out. Finally, we reconstruct the dynamic scene parts such as humans and map them into the static scene using the existing geometric information to facilitate temporal referencing of the data.

The subsequent visual analytics module combines the static and dynamic reconstructed scene parts and automatic detections and visualizes them in a single user interface. Users can spatially (3D) and temporarily navigate the scene, allowing them to inspect the scene from any angle and time where footage is available. This more intuitive way of navigating through many different videos recorded by various cameras in a scene spares the investigator much time viewing the video footage manually. Automatic detections support the investigator to find meaningful aspects in space and time. However, automatic detections sometimes miss important events. Therefore, we do not only rely on detections but augment them into the reconstructed scene such that the user can always inspect the original video footage.

2 Related Work

This section focusses on known solutions on 4D dynamic (crime) scene reconstruction and visualization.

2.1 Dynamic Scene Reconstruction

Dynamic objects not only are difficult to reconstruct using off-the-shelf Simultaneous Localization and Mapping (SLAM) or Structure from Motion (SfM) algorithms, but also impede reconstruction quality for static scenes. Zhong et al. [19] present a joint Detect-SLAM framework that simultaneously addresses the reconstruction and detection problem for dynamic objects, which improves in both cases the reconstruction quality and detection rate under unusual viewpoints and occlusion by first

segmenting the scene. Similar to this method, Bullinger et al. [3] employ segmentation techniques in conjunction with the optical flow to obtain object-specific motion cues and corresponding points. Then, SfM and triangulation can be applied to enable 3D reconstruction and tracking for static and dynamic scene parts, respectively. However, this approach presumes the use of stereo cameras, which are not always available.

Another relevant research direction is the direct recovery of 3D dynamic scenes, which is more challenging and demands more elaborate routines to mitigate the negative influences in dynamic scenarios. Ji et al. [10] focus on the reconstruction of a single dynamic foreground subject and addresses video synchronization by exploiting locally rigid patches without the need for segmentation. On the contrary, Mustafa et al. [13, 12] propose to improve on an initial sparse reconstruction using classic reconstruction techniques with a joint optimization framework. They take data, contrast, smoothness, as well as temporal terms into account to constrain the solution space to get a clean depth recovery for synchronized and unsynchronized input videos. However, the algorithms focus on few large moving foreground objects in the scene, which is difficult to adapt to real-world scenarios, where wide areas of many small dynamic objects, such as persons and cars, are present.

2.2 Crime Scene Reconstruction and Visualization in 4D

3D laser scans have seen increased prevalence in crime scene investigations across the globe thanks to the visual context and accuracy compared to 2D pictures, as well as the continuous lowering of device costs [14]. 3D photogrammetric sensing techniques, such as SfM, offer more advantages by the easiness of capturing 2D images, and ubiquity of mobile phone cameras nowadays. Despite the recent strides achieved in the machine learning and computer vision societies, real-world applications of 3D/4D crime scene investigation systems are rare.

Baier and Rando utilize SfM to improve mass grave documentation in archaeological investigations [1]. Urbanová et al. [18] and Michienzi et al. [11] test commercial 3D reconstruction solutions for the recovery of 3D human body surfaces in forensic pathology and injury documentation. All studies demonstrate higher flexibility and lower time cost compared to 3D laser scanners with comparable accuracy in most of the cases. This provides new perspectives for high-precision static 3D reconstruction.

In a current and ongoing project [5], SfM is adopted to temporally and spatially align vast live footage from cameras and smartphones by eyewitnesses during London's Grenfell Tower fire in 2017 that took 72 lives. After motion tracking of the fire and stabilization of the videos, the image frames are projected onto the wire-frame model of Grenfell Tower, such that the catastrophe can be unfolded and explored in 4D.

Bostanci showcases a more proper setup, in which the author developed an interactive investigation tool for 3D reconstruction of crime scenes [2]. An off-the-shelf SfM algorithm Bundler [17] with standard workflows for sparse 3D structures and CMVS [6] for dense reconstruction are leveraged. Manual registration is needed to initialize the merging of point clouds from different clusters. Furthermore, interactive measurement operations for distance calculation is also included. As opposed to the presented tool in this paper, the software in [2] offers minimal and basic functionalities, lacking the capability for reconstructing dynamic objects, geo-registration and progressive expansion of reconstructions in an automated fashion.

Finally, we refer the reader to [4] for a comprehensive review of image-based modalities for forensic investigations.

3 Crime Scene Reconstruction

Our crime scene reconstruction approach is divided into four stages. The first stage is the pre-processing stage, which is done primarily to save computational cost for large data sets. During this stage, videos are sampled into one multiple image files depending on whether the content originates from static or moving cameras. Binary masks are computed for every frame to differentiate between dynamic scene parts like persons or cars and the static scene which improves the reconstruction quality. In the second stage, a static 3D scene reconstruction is computed based on an SfM approach. This results in poses with six degrees of freedom (6DOF) for ideally all cameras, encoding the cameras position and viewing direction. Furthermore, a dense point cloud and a surface mesh are computed, which are useful to provide investigators context for the scene in which an incident has happened. In the third stage, we perform a geo-registration of the scene using satellite imagery, recovering the absolute scale, and enable distance measurements in meters. At this stage, it is also possible to manually set the 6DOF camera pose of images that were not automatically reconstructed, since an automatic registration of some images may have failed that are heavily occluded by dynamic objects in crowded scenarios. Finally, in the last stage, a dynamic scene reconstruction is performed. Depth maps are estimated for all video frame from moving or static cameras and are embedded into the static scene. This gives investigators an impression of a temporal 3D reconstruction, also known as 4D reconstruction, that can be played like a video where the user can additionally freely navigate in space and time. The resulting 4D scene reconstruction can then be analyzed with the crime scene visualization tool that is described in Section 4.

3.1 Pre-processing

A significant problem when reconstructing a 4D scene from video footage is the long wait time due to the processing of mass data. All video footage is manually annotated in the first

Figure 1. The binary mask to separate static and dynamic scene parts based on Mask R-CNN.

stage to reduce the wait time in our approach. The user can provide information on whether footage originates from a static surveillance camera or a mobile device. This labeling is useful to save computational cost by excluding frames that are not necessary. For example, not every single camera frame from a static camera has to be processed when estimating its 6DOF pose and also not every single frame adds new information to the reconstruction of the static scene. Furthermore as shown in Figure 1, binary masks are automatically computed using Mask R-CNN [9] for all frames, which allows the pixel-wise differentiation between dynamic and static scene parts to perform the static scene reconstruction only on the static image regions.

3.2 Static Scene Reconstruction

The reconstruction of the static 3D scene is based on a popular state-of-the-art SfM approach [15, 16], which allows the full automatic computation of large scale 3D scene reconstructions from arbitrary image collection with overlapping views by matching SIFT features between all camera views. We additionally make use of binary masks to prevent matching SIFT descriptors of potentially dynamic scene parts, like persons, to prevent inconsistencies in the reconstruction. The result is a set of 6DOF camera poses in a common coordinate space and intrinsic camera parameters. Based on these camera poses, a dense point cloud and a surface mesh are computed to represent the static scene. Figure 2 demonstrates the result of an example reconstruction containing three surveillance camera views and footage from two mobile cameras. The reconstruction is not complete (see white holes), which is because the videos used for reconstruction did not completely cover the scene. However, this can be addressed by adding more image or video data to the reconstruction pipeline.

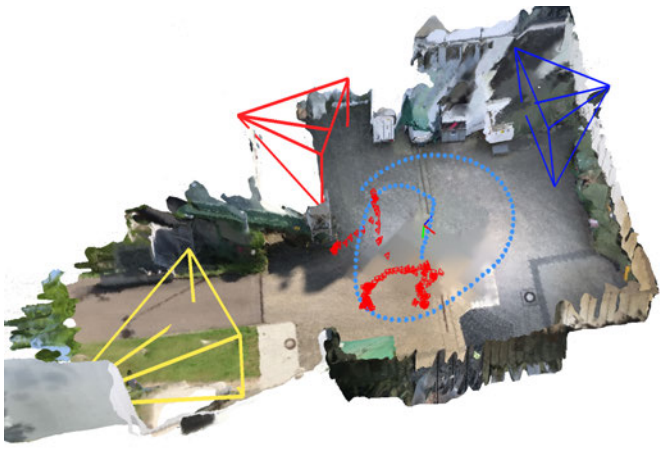

Figure 2. The result of the SfM-based static scene reconstruction. Surveillance cameras are shown as large fields of views. The blue smaller field of views represents camera poses from images of a video that has been taken after a crime scene incident and re red field of view depict camera poses of footage that has been recorded by a witness during an event.

3.3 Geo-registration

SfM algorithms typically do not recover an absolute scale in meters, meaning that distances cannot be directly measured from a plain 3D reconstruction if no further metadata is provided, like GPS locations in EXIF metadata. A problem here is that video data typically does not provide EXIF metadata and GPS tagged images may suffer from limited positional accuracy. However, to address the scale problem, we provide a tool that allows to manually geo-register a scene reconstruction from 3D-3D correspondences between 3D points of the reconstruction and 3D points on a world map that consists of satellite imagery with an elevation map. A set of four 3D-3D correspondences is sufficient; however, providing more correspondences allows for a more robust registration, as a RANSAC-based registration algorithm with an outlier detection can be used. An advantage is that by selecting just a few correspondences from a few image views, automatically all cameras locations and object points are represented in a global UTM or GPS coordinates. The geo-registration as well solves the arbitrary orientation of the reconstruction as it aligns the reconstruction to the map. Furthermore, after a successful geo-registration, information about the ground plane and gravity vector is available. This can be used as metadata for the reconstruction of the dynamic scene parts like persons that are moving through the scene.

3.4 Dynamic Scene Reconstruction

The reconstruction of the dynamic scene parts is crucial for enabling 4D exploration of a crime scene. The 6DOF camera poses and intrinsic camera parameters are already recovered in the static 3D scene reconstruction. A depth map has to be calculated for each video frame to add the dynamic dimension. Classical approaches [8] make use of a stereo camera consisting of synchronized cameras or RGBD cameras and compute a disparity map that implicitly encodes depth. However, in real-world scenarios, there are typically no synchronized cameras which additionally fulfill the requirement of sufficient overlaps to allow for depth estimation. The fact that a lot of data has to be processed makes approaches infeasible that exceed a con-

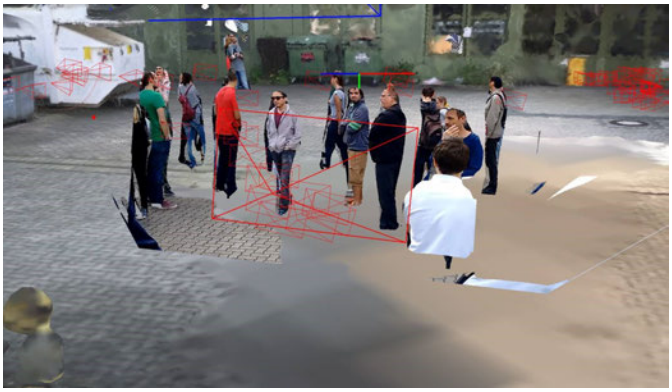

Figure 3. An embedded dynamic reconstruction in the static 3D reconstruction for a single time point t.

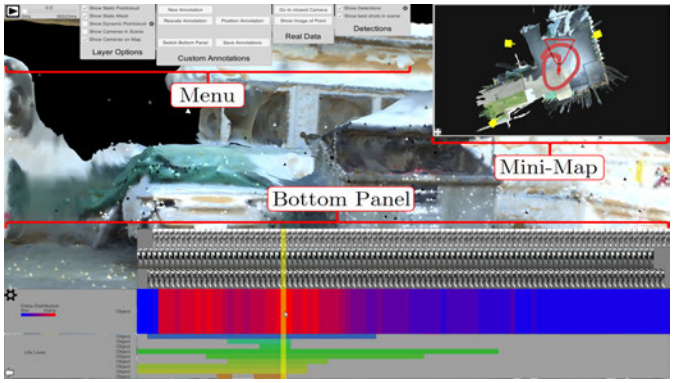

Figure 4. The visual exploration interface of the 4D reconstruction, comprising three components placed at the edges of the reconstruction view: menu-bar, mini-map, and time-line.

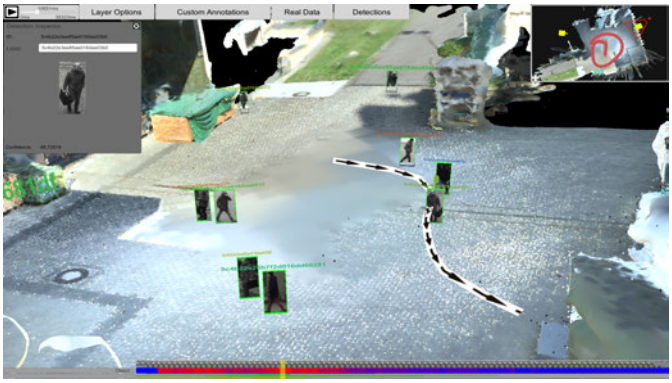

Figure 5. Automatic detections from all cameras are embedded in the 3D scene. Detections can be selected to blend in its track over time. Additionally, an information panel is shown (top-left) that depicts, e.g., the best-shot of an entity.

stant overhead by considering multiple frames for computing a single depth map. Recently there has been intensive research on monocular depth estimation using convolutional neural networks [7], allowing to estimate a depth map for a given input image. However, the problem is that these networks are typically learned for very specific scenes targeting to replace specific stereo camera setups. Since the depth maps are very noisy, they make the analysis rather tricky for investigators. In our approach, we created a monocular depth estimation approach for persons, making use of Mask R-CNN, a neural network, that creates a per person instance segmentation as well as 2D pose keypoint estimations. Given the camera location, the intrinsic camera parameters, and the ground plane equation, we estimate each person's location in space and map the depth values of each instance segmentation to a plane that is orthogonal to the ground plane, while all other pixels are mapped to the ground plane. A disadvantage of this approach is that all persons are mapped as flat objects. However, an advantage is that a flat depth map is better for investigators as the depth map does not have any noise, while the results look nearly photo-realistic. Figure 3 shows an example embedding such a dynamic monocular reconstruction into the static reconstruction. Figure 7 shows another dynamic reconstruction in the visual exploration interface presented in this paper.

4 Crime Scene Visualization

This section focuses on the visualization of the 4D reconstruction. The visualization is enriched with automatic object and person detections from the original video footage. The following details the interface and shows how it can be used to visually trace a sequence of events from an incident that is subject to investigation. The interface is explained by presenting an exemplary event that has been recorded with four cameras. Three of the cameras are static surveillance cameras, and the last is a mobile phone camera that was used to record the environment after the incident took place.

4.1 Interface

The interface for the exploration of the 4D scene consists of four components. The main area depicts the reconstructed 3D scene and is surrounded by three other panels that are shown in Figure 4: a menu-bar at the top-left, a mini-map at the top-right, and a timeline at the bottom. The menu-bar provides the functionality to configure the content that is displayed in the scene. For instance, users can choose whether the camera positions or how automatic detections should be displayed within the scene. The static reconstruction of the environment is created by using a video recording from a moving device that covers the entire scene. In this case, we used a mobile phone camera and walked through the scene, as depicted with red camera icons showed inside the mini-map in Figure 4.

As shown in the top-right mini-map, the scene was recorded by three static surveillance cameras (marked by yellow color). By clicking on a yellow camera icon, the user can change the viewport to the respective position and inspect the scene from the camera's location. Additionally, it is possible to view the original camera footage of the respective camera. In this way, we put all available video footage of the incident into a common context and facilitate the identification of which cameras monitored different area at different points in time.

Automatic detections found by machine learning algorithms are provided by project partners that we embed in the scene (see Figure 5). For each footage video, minimal bounding rectangles (MBRs) are extracted for objects of interest, such as persons, as a preprocessing step and saved using image coordinates. Their locations are determined by using raycasting and the 3D model of the scene. Paths represent MBRs of the same person or object in successive frames. For each path, we are provided with a best-shot that displays the tracked entity optimally (e.g., a frontal shot of a person with high resolution). For the display of detections in the 3D scene, we provide three different options (see Figure 6): (i) the MBR itself; (ii) the best-shot; (iii) the actual image of the respective frame and MBR. Upon selection, the respective path is visualized, as shown in Figure 5.

Figure 6. Automatic detections can be displayed as MBRs (left), snippets from video footage (center) or the best-shot of the entity from the entire scene (right).

Besides the display of a static 3D environment with automatic detections, we enable the display of the video footage as 3D point clouds. For each camera, depth-maps are generated (see Section 3.4). For each camera, each pixel is projected into the scene resembling the video footage as 3D point clouds. We use the position and orientation of the cameras (extrinsic) in the scene and their camera intrinsics to place the pixels at the correct scene locations based on the predefined depths of each pixel using given depth maps.

4.2 Visual Exploration of 4D Crime Scene

The user can navigate the 4D scene spatially and temporally. The keyboard can be used for spatial navigation. A time-slider at the top-left, as well as a timeline at the bottom, can be used to investigate time frames of interest (see Figure 4). Besides selecting single times for detailed inspection, the user can play the scene and view the progression of events. Detections from all cameras are placed into a common context and can be followed simultaneously. By hovering the top row of the bottom panel (see Figure 8), previews of frames are enlarged (fish-eye effect), allowing the user to browse the entire timeline for potentially interesting segments quickly.

To provide an overview of automatically extracted detections, we created a heat-map visualization in the bottom panel. The user can filter by detection types and visualize the distribution of the selected detections over all frames. The heat-map

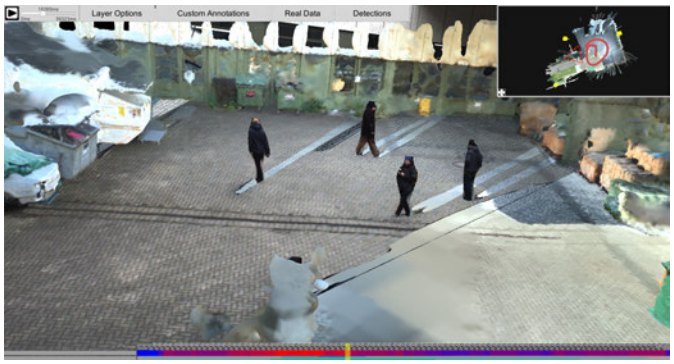

Figure 7. A dynamic point cloud from a mono-camera. For each frame in the input footage, a depth map is estimated. From the origin of the camera in the 3D scene, we project each input frame pixel into the scene respecting their depth information.

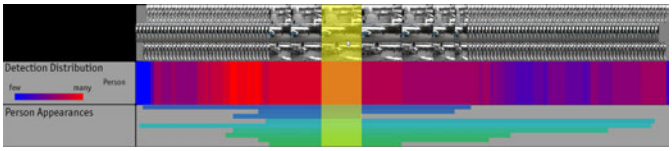

Figure 8. Bottom panel visualizations. The top element serves as a frame preview. In the center, a heat-map depicts where in the timeline many entities of the selected class (here "Person") were detected. At the bottom, lines indicate, for each detection in the current frame, when it first appeared in the scene and when it disappeared again.

visualization depicts at what time many detections occurred. In particular, for large video sequences, this can be useful to detect areas of interest quickly. For instance, in a video of 10 hours, sequences with suitcases can be identified in seconds.

Below, lifelines illustrate at what point of time an entity enters and leaves the scene (Figure 8, bottom). This visualization can be used to jump to the entry or exit point of an object of interest. For instance, when the user recognizes a suitcase standing on the floor at a specific time-frame in the video, the point in time when it was placed, there can be identified quickly.

Additionally, the user can create custom annotations, position them in the 3D scene, and animate them to change their positions over time (see Figure 9). Manual annotations are useful for embedding details from witness reports in the given evidence context. Testimonies can be confirmed or rejected by checking if the given circumstances are even possible (e.g., "if person X walked at time Z from A to B - was it possible for him to see person Y?").

We experiment with virtual reality environments (VREs) to provide an immersive experience to the investigator. The user enters the 4D scene with a head-mounted display and can walk naturally around in it. The 3D environment is perceived in familiar stereoscopic 3D, and real-world distances apply. Within a VRE, the user can better estimate distances as the scene is scaled relatively to the user by the geo-registration step.

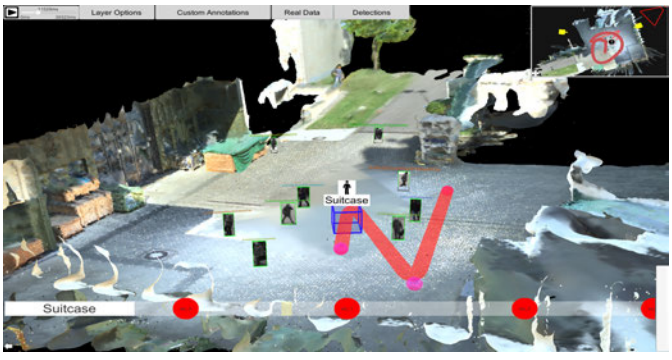

Figure 9. Manual annotations. For instance, based on witness reports, the movement of a suspicious person with a suitcase is visualized within the context of automated detections.

5 Conclusions

We presented a novel 4D scene reconstruction and visualization framework for mass video data, which is, to the best of our knowledge, the first tool for dynamic crime scene investigation. Based on recent advances in computer vision and machine learning, state-of-the-art deep-learning-based segmentation and human pose estimation approaches are leveraged to reconstruct and register dynamic objects into the geo-registered 3D static scene model. Moreover, a visualization frontend is developed to allow for intuitive visualization of the 4D crime scene with plenty of exploration possibilities. We support analysts to interactively incorporate information from a criminal case such as witness statements. The user can add annotations that can describe the movement of objects or persons. Embedding such a vast amount of potentially heterogeneous information into a single environment supports the mental model of analysts. Therefore, decision-makers can simultaneously visualize all available related contextual information and are supported in making time-critical decisions. Ongoing evaluations and liaisons with end-users reveal application areas beyond the intended crime scene reconstruction such as live monitoring of public areas (e.g., airports and train stations) and using reconstructions for tactical training and analysis of police forces.

6 Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation program in the context of the VICTORIA project under grant agreement No. 740754.

References

[1] W. Baier and C. Rando. Developing the use of structure-from-motion in mass grave documentation. *Forensic Science International*, 261:19–25, 2016.

[2] E. Bostanci. 3D reconstruction of crime scenes and design considerations for an interactive investigation tool. *International Journal of Information Security Science*, 4(2):50–58, 2015.

[3] S. Bullinger, C. Bodensteiner, and M. Arens. 3D object trajectory reconstruction using stereo matching and instance flow based multiple object tracking. In *International Conference on Machine Vision Applications (MVA)*, 2019.

[4] R. M. Carew and D. Errickson. Imaging in forensic science: Five years on. *Journal of Forensic Radiology and Imaging*, 16:24–33, 2019.

[5] Forensic Architecture. The Grenfell tower fire. https://forensic-architecture.org/investigation/the-grenfell-tower-fire. Accessed: 2019-08-12.

[6] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[7] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016.

[8] R. Hamzah, R. Abd Rahim, and Z. M. Noh. Sum of absolute differences algorithm in stereo correspondence problem for stereo matching in computer vision application. 1:652–657, 2010.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[10] D. Ji, E. Dunn, and J.-M. Frahm. Spatio-temporally consistent correspondence for dense dynamic scene modeling. In *European Conference on Computer Vision (ECCV)*, pages 3–18, 2016.

[11] R. Michienzi, S. Meier, L. C. Ebert, R. M. Martinez, and T. Sieberth. Comparison of forensic photo-documentation to a photogrammetric solution using the multi-camera system "Botscan". *Forensic Science International*, 288:46–52, 2018.

[12] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5583–5592, 2017.

[13] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton. General dynamic scene reconstruction from multiple view video. In *IEEE International Conference on Computer Vision (ICCV)*, pages 900–908, 2015.

[14] D. Raneri. Enhancing forensic investigation through the use of modern three-dimensional (3d) imaging technologies for crime scene reconstruction. *Australian Journal of Forensic Sciences*, 50(6):697–707, 2018.

[15] J. L. Schönberger. COLMAP – SfM and MVS. `https://demuc.de/colmap/`, 2018.

[16] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics (TOG)*, 25(3):835–846, 2006.

[18] P. Urbanová, P. Hejna, and M. Jurda. Testing photogrammetry-based techniques for three-dimensional surface documentation in forensic pathology. *Forensic Science International*, 250:77–86, 2015.

[19] F. Zhong, S. Wang, Z. Zhang, C. Chen, and Y. Wang. Detect-SLAM: Making object detection and SLAM mutually beneficial. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1001–1010, 2018.