CO-REGISTRATION OF VIDEO-GRAMMETRIC POINT CLOUDS WITH BIM – FIRST CONCEPTUAL RESULTS

T. Kaiser[1], C. Clemen[1], M. Block-Berlitz[2]

[1]HTW Dresden, Faculty of Spatial Information, Dresden, Germany - tim.kaiser, christian.clemen@htw-dresden.de
[2]HTW Dresden, Faculty of Informatics/Mathematics, Dresden, Germany - marco.block-berlitz@htw-dresden.de

Commission V, WG V/7

KEY WORDS: BIM, Videogrammetry, Registration, SfM

ABSTRACT:

The co-registration of photogrammetric products such as image blocks or point clouds is an essential step before they can be used for subsequent analysis. Usually this is done by using control points. This has some disadvantages such as the need for additional measuring devices and a laborious measuring of the coordinates. In prior works we developed a procedure that enables a marker-less co-registration of an image block with a digital building model. This extended abstract presents our current research as work-progress. For further facilitating and improving this process we identified two tasks. Using videogrammetry as data capturing technique and using an enhanced matching algorithm during the co-registration. This paper summarizes essential steps when making the switch from photogrammetry to videogrammetry and explains the basic principles of the improved matching process.

1. INTRODUCTION

Today the sustainable maintenance and conservation of buildings and especially infrastructure such as bridges, tunnels and roads is a major challenge. New tools for the digital documentation of the actual conditions can help to detect necessary renovation measures on time. Photogrammetric measuring techniques can help to improve this process. Point clouds and oriented image blocks can be used for capturing the actual state of the structure for different points in time and therefore to monitor the health of it. Modern photogrammetric sensoros providing a lot of details in a high resolution combined with artificial intelligence techniques can for example be used to detect cracks or deformations on the structure (Morgenthal et al., 2021).

An important step in the photogrammetric process chain is the registration of the generated data. A proper registration either in respect to a global reference frame or a digital model is very important to establish the connection of potential damages and their location in the structure. Usually the registration is carried out using control points with known coordinates in the world as well as in the object coordinate system. This well established procedure has the advantage that high registration accuracies can be reached. On the other hand it often requires additional measuring devices such as total stations or GNSS receivers for obtaining the coordinates of the control points. Additionally, this requires expert knowledge and manually measuring the control points in the images is an error-prone, repetitive and time-consuming task.

In order to get widely adopted by many users it is important that the complete process including data capturing and the actual co-registration can be automated as high as possible. With the emergence of Structure from Motion (SfM) packages it already became possible to reconstruct accurate 3d scenes if certain conditions such as enough overlap between the images are met. For further simplifying the data capturing, video frames can be used as input data source.

In (Kaiser et al., 2022) we presented a novel approach for the automated co-registration of (single) image blocks with an existing digital building model. With our ongoing research we want to improve and ease the complete workflow by using videogrammetry as data capturing technique (Section 4) and an enhanced matching algorithm (Section 5). Section 4 discusses the various principles of image selection from video frames in the context of videogrammetric 3d reconstruction. Also the videoprocessing pipeline is presented. Section 5 prsents a new principle component based cluster method for the SfM-generated 3d-lines. This method serves to reduce the of candidates and intends do accelerate the matching algorithms from image blocks to BIM model. Please note, that the two enhancements are theoretically independent, but are practically used in a common pipeline for the co-Registration of videogrammetric point clouds with BIM.

2. RELATED WORK

The co-registration of photogrammetric products with digital building models is a very active research field. The rising usage of digital methods like Building Information Modeling (BIM) has accelerated this trend. In projects related to construction progress monitoring (Vincke and Vergauwen, 2020, Tuttas et al., 2017) the registration is carried out once at the begin of the construction using a classical approach with control points. Image blocks of later points in time are then co-registered with the initial reference frame.

Other works use the geometry of the digital building model for an automated co-registration. (Kim et al., 2013) for example co-registers a point cloud to a model with the help of the Iterative Closest Point Algorithm. (Kropp et al., 2018) match lines extracted from video sequences with lines extracted from a building model for co-registering the image block. Whereas plane-based registration mainly is used in applications related to terrestrial laser scanning (TLS). These procedures can either be used to register the single scan stations into one common

reference frame (Wujanz et al., 2018) or to co-register the scan with a building model (Bosché, 2012).

3. EXISTING SOLUTION

As stated in the introduction, we developed a procedure that enables the co-registration of an image block consisting of single images with a digital building model (Kaiser et al., 2022). More precisely we focused on the co-registration of indoor scenes. The basic idea of the method is to match 3d line segments that are extracted from the images with planar surfaces from the digital building model. By observing the geometric relationships between lines and planes the required transformation parameters can be estimated. Figure 1 shows the basic steps for co-registering an image block with the building model.

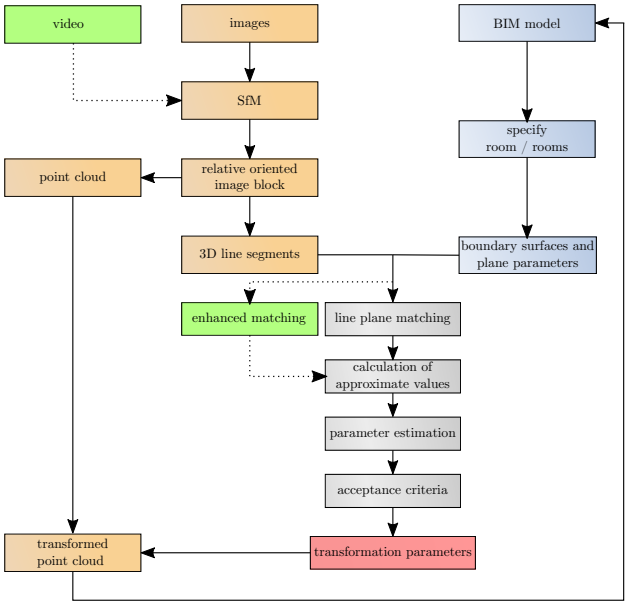

Figure 1. Workflow of existing solution with BIM (blue), photogrammetry (orange), co-registration (grey), result (red) and the planned extensions (green). Only the new video data source and the enhanced matching are discussed in this paper.

After the images have been captured they are relatively oriented using Structure from Motion (SfM) algorithms. In our implementation this is done using the open source software *COLMAP* (Schönberger and Frahm, 2016). This step delivers the interior and exterior orientation of the images. The orientation parameters and the images are processed by *Line3D++* (Hofer et al., 2017) for extracting the 3d line segments. These are defined by the coordinates of the start and end points in the SfM coordinate system

The necessary planar surfaces are extracted from a digital building modeled with the open Industry Foundation Classes (IFC) standard. Especially for indoor spaces IFC implements the concept of *Space Boundaries* that define the individual room bounding surfaces [1]. For planar surfaces with the normal vector $\overrightarrow{n} = [abc]^T$ a plane equation can be formulated:

$$ax + by + cz - d = 0 \tag{1}$$

[1] `https://technical.buildingsmart.org/standards/ifc/ifc-schema-specifications/`

The required transformation parameters consisting of

- three rotations around the coordinate axes, equal to the rotation Matrix R,
- three components of the translation vector $\overrightarrow{t}$,
- s scale parameter m

are determined in the adjustment stage. Each 3d line segment that is directly located on an extracted boundary surface provides two observation equations

$$f\left(x, l + v\right) = \langle R \cdot \vec{u}, \vec{n} \rangle = 0 \tag{2}$$

and

$$f\left(x, l + v\right) = \left\langle \left(m \cdot R \cdot \vec{p} + \vec{t}\right), \vec{n} \right\rangle - d = 0 \tag{3}$$

where R is the rotation matrix, $\overrightarrow{t}$ the translation vector, m the scale parameter, $\overrightarrow{n}$ is the normal vector of the plane, $\overrightarrow{u}$ is the direction vector of the 3d line segment and $\overrightarrow{p}$ is the mid point of line.

Equation 2 can be used to calculate the unknown rotation from the point cloud to the BIM coordinate system whereas equation 3 also enables to determine the translation and the scale parameters. By using a Gauß Helmert Model the transformation parameter can be estimated.

The adjustment process only delivers correct results if the involved 3d line segments are matched to the correct planar surface. However, there is no a priori information about correct line plane pairs available. Since a brute force approach (where all possible combinations of line plane pairs would be tested) is not feasible, we developed a clustering algorithm that assigned the 3d line segments into multiple clusters. In the next step a RANSAC (Fischler and Bolles, 1981) inspired random assignment of the cluster's lines to the planes is performed. In total four line plane pairs are necessary to calculate the transformation parameters. Due to the random line plane assignment, numerous minimal configurations have to be processed and afterwards filtered to obtain the best suitable seven transformation parameters $R, \vec{t}$ and m.

4. VIDEOGRAMMETRY

In recent years, and by now decades, development in the field of real-time robotics has come a long way in terms of camera-based systems. In just a few milliseconds, vehicles can recognize signs and road situations and, in some cases, react autonomously to them. A very active research focus in this context uses SfM and pursues Simultaneous Location and Mapping (SLAM) solutions to localize themselves in self-generated maps or 3d models of an environmental situation. The boundaries between these real-time applications and photogrammetry methods are now fluid. Both profit greatly from this. Videogrammetry (VG) can be understood simplified as an extension of photogrammetry (PG) by an intelligent image selection (IS) in the available videos:

$$\text{VG} = \text{IS} + \text{PG} \tag{4}$$

The approach of capturing and processing videos instead of photos has a number of advantages and disadvantages. The

biggest disadvantage is certainly the fact that the extracted single photos usually do not have geotags and thus an automatic georeferencing is not easily possible. However, this disadvantage can be solved satisfactorily in combination with pure photogrammetry. For the image selection we have many different strategies at our disposal. First of all: Not only one solution exists for the image selection. If the goal is to generate a point cloud as quickly as possible, e.g. to ensure on site that the acquired data is complete and to generate a coherent, gapless 3d model, then a minimal, fast image selection would be an option. However, if the goal is to generate the densest point cloud possible, then more time can be invested in image selection and that may result in a larger image set. Before discussing some of these strategies, we need to understand the spectrum of data and the min-max conflict that exists with it.

4.1 Min-Max Conflict

For 3d reconstruction of a point in SfM, there must be at least three images in which that point has been uniquely determined. Since we have a continuum of consecutive data available in the video footage, we could come up with the idea of just taking all the frames. This would give us the minimum distance between frames. Given the rule of three, we can derive the min property: The smaller the distance between the images, the more 3d points are possible. In practice, we quickly find that using all the images unfortunately leads to worse results with fewer 3d points than using a smaller number of images. In order to understand the reason, we need to appreciate another important property in the SfM approach: The larger the distance between images, the more accurately 3d points can be determined, and only accurate 3d points survive later in the 3d model (see Figure 2).

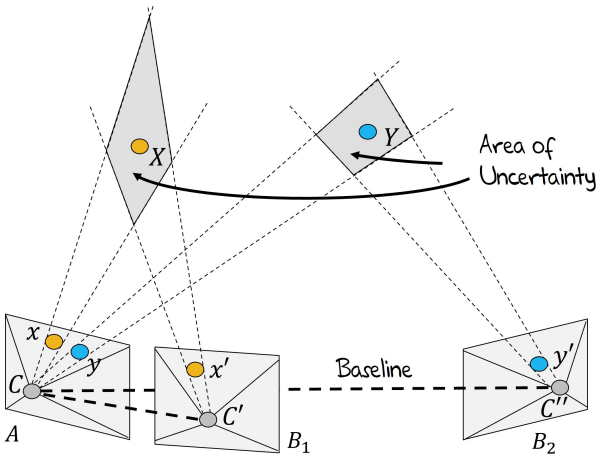

Figure 2. Relation between baseline length and the resulting area of uncertainty when triangulating a point.

If for two images A and B_1 the image distance of the camera centers (baseline) is smaller, we get a larger area of uncertainty for the jointly observed point X than if we choose a larger image distance as for images A and B_2. The area of uncertainty provides us an important quality attribute for the identified 3d point. Thus, in order to obtain the maximum number of 3d points for a 3d model, we need to solve the so-called min-max conflict during image selection: maximize 3d points by choosing an appropriate image (image spacing) between min and max property.

4.2 Image and correspondence evaluation criteria

The research in feature extraction, the reduction of all image pixels to some relevant, is as old as the Computer vision itself. There are several solutions i.e. Harris Corner (Harris and Stephens, 1988), SIFT (Lowe, 2004) or SURF (Bay et al., 2006), and much more. All candidates for SfM need to be invariant to affine transformations like scaling, rotating, translation and a mix of them. One of the most common feature detectors and used in various applications like object recognition, image retrieval or 3d reconstruction is SIFT, published and patented by Lowe (2004). There are different approaches to speed them up, like SiftGPU (Wu, 2010). But in robotics, when Computer vision needs to work in real-time, other solutions are more common (Miksik and Mikolajczyk, 2012).

Typically, a feature extracted by a detector has not only a position. In most cases, i.e. to improve the necessary matching between two feature sets extracted from two images I_1 and I_2, every feature has a more accurate description (Mikolajczyk and Schmid, 2005). We write a set of image keypoints as $K_{x,y}(I)$ extracted with detector $x \in \{SIFT, SURF, \ldots\}$ and descriptor $y \in \{SIFT, GLOH, \ldots\}$ in image I and $\|K_{x,y}(I)\|$ counts the number of detected features. If there are two images I_i and I_j with feature sets $K_{x,y}(I_i)$ and $K_{x,y}(I_j)$ then $K_{x,y}(I_i, I_j)$ is the set of corresponding keypoints for the image pair using RANSAC (Hartley and Zisserman, 2011). Appropriate the number of features is $\|K_{x,y}(I_i, I_j)\|$. To quantify the relative intersection set in percent, we notice $[\![K_{x,y}(I_i, I_j)]\!]$ with:

$$[\![K_{x,y}(I_i, I_j)]\!] = \frac{\|K_{x,y}(I_i, I_j)\|}{max\left(\|K_{x,y}(I_i)\|, \|K_{x,y}(I_j)\|\right)} \quad (5)$$

It is easily comprehensive, that the result for $[\![K_{x,y}(I, I)]\!]$ should be always 1. The following sections describe methods to measure the quality of single images or the quality of correspondences between two images, considering the aim which is to use them in a 3d reconstruction process. Most of these methods are based on feature detection. If e.g. a method $A(I_i, I_j)$ is given to deliver a correlation score between images I_i and I_j which followed by the step $A(K_{x,y}(I_i), K_{x,y}(I_j))$ we simplify to $A(I_i, I_j)_{x,y}$. If selecting one image as the n^{th} keyframe from a set of $\{I_1, I_2, \ldots, I_{max}\}$, we notice this image $\mathbf{I^n}$ by bold letters. If the position i inside the image set is needed to follow the algorithm, we give both indices $\mathbf{I_i^n}$.

A number of solutions are available today for solving the image selection part due to the Min-Max Conflict in real-time scenarios, let's take a look at a few representatives of these algorithms.

4.2.1 Sharpness measure While recording video data with moving systems like UAVs, single images with the same content may differ strongly in sharpness due to the fact that small camera movements are applied. In contrast to the relative sharpness measure for an image I, as a mean square of the horizontal and vertical derivatives:

$$S(I) = \frac{1}{2\,\|I\|} \int\int \left(\frac{\partial I(x,y)}{\partial x}\right)^2 + \left(\frac{\partial I(x,y)}{\partial y}\right)^2 dxdy \quad (6)$$

Nistér proposes a discretized, faster version with finite differences except for the image boundaries

$$S(f,I)=\frac{1}{2\|I\|}\sum_{(x,y)\in I}(f(x+1,y)-f(x-1,y))^2+(f(x,y+1)-f(x,y-1))^2, \tag{7}$$

where $\|I\|$ conforms to the amount of pixels and f describes an image function to get pixels from downsampled and normalized image data (Nistér, 2001).

4.2.2 Normalized Correlation Constraint Nistér uses the normalized correlation constraint

$$NC(I_i, I_j) = [\![K_{x,y}(I_i, I_j)]\!] \tag{8}$$

between two images I_i and I_j to delete redundant frames (Nistér, 2001). Redundant in that case means very similar and will be discussed later.

4.2.3 Distance Constraint Nistér also checked the maximum distance in correspondences (Nistér, 2001). The estimated homography mapping H does not violate the maximum expected disparity d at any point.

In combination with the Normalized Correlation Constraint (Sec. 4.2.2), he deletes frame I_i when maximum distance is smaller than T_d with $T_d = \frac{image\,size}{10}$.

4.2.4 Correspondence Ratio Constraint Seo et. al. (Seo, 2008) considered $\|K_{x,y}(I_i, I_j)\|$, the ratio of the number of correspondences to the total number of features, for correspondence:

$$CRC(I_i, I_j)_{x,y} = \frac{\|K_{x,y}(I_i, I_j)\|}{\|K_{x,y}(I_i)\|} \tag{9}$$

The Correspondence Ratio Constraint CRC depends on the camera motion and needs to be located between the values t_{low} and t_{high}, which are not specified by the authors. Rashidi et. al. experimented with scenes of different complexity and different camera motion speed and suggested estimated values for them (Rashidi et al., 2013).

4.2.5 Maximum Distance Constraint A simple method motivated by autonomous robot navigation and proposed by Royer et. al. selects images with maximum distances while there are at least M common interest points between two correlated frames (Royer et al., 2007). They choose always the first image as the first keyframe $\mathbf{I_1^1}$. When n keyframes $\mathbf{I^1}, \mathbf{I^2}, \ldots, \mathbf{I^n}$ are chosen, they select the next keyframe $\mathbf{I^{n+1}}$ as follows: (i) there are as many video frames as possible between $\mathbf{I^n}$ and $\mathbf{I^{n+1}}$, (ii) there are at least M interest points in common between $\mathbf{I^{n+1}}$ and $\mathbf{I^n}$, (iii) there are at least N common points between $\mathbf{I^{n+1}}$ and $\mathbf{I^{n-1}}$.

We can summarize this description as follows

$$\begin{aligned} &MDC(\mathbf{I^{n-1}}, \mathbf{I_i^n}, [I_{i+1}, \ldots, I_{max}])_{x,y} = \\ &\max_k(\|K_{x,y}(\mathbf{I^n}, I_k)\| > M, \left\|K_{x,y}(\mathbf{I^{n-1}}, I_k)\right\| > N), \end{aligned} \tag{10}$$

where $i < k$. The two unknown parameters M and N are specified by the authors with $M = 400$ and $N = 300$ and were set experimentally. Royer et. al. (2007) use Harris corner detector for feature detection.

4.2.6 Optical-Flow-Based Motion Estimation In 2001 Nistér uses the initial step of coarse to fine optical flow based video mosaicing (Kanatani and Ohta, 1999) to use the result as a global motion estimation for Structure and Motion (Nistér, 2001). The motivations to use this over feature based approaches like in Capel et. al. (Capel and Zisserman, 1998) were that the behavior works fast and also for gravely unsharp frames. Assuming a rigid world, between two images I_1 and I_2 a homographic mapping H can be derived.

An image I_i is downsampled and normalized and a position of pixel $\overrightarrow{p} = (x, y)$ is accessible by an image function $f_i(\overrightarrow{p})$ (see Sec. 4.2.1). To estimate H the mean square residual R with

$$R(f_1, f_2, H, \vartheta) = \frac{1}{\|\vartheta\|}\sum_{x\in\vartheta}(f_2(H\overrightarrow{p}) - f_1(\overrightarrow{p}))^2 \tag{11}$$

will be minimized using non-linear least squares algorithm such as Levenberg-Marquardt (Press et al., 1988).

4.2.7 Degeneracy Constraint As the fundamental matrix F better defines general camera motion, the homography H better defines degenerated camera movements. The Geometric Robust Information Criterion $GRIC$ introduced by Torr

$$GRIC(I_i, I_j)_{x,y} = \sum_i p(e_i^2) + \lambda_1 dn + \lambda_2 k \tag{12}$$

computes a score based on the fundamental matrix $GRIC_F$ and the homography $GRIC_H$ separately (Torr, 1998), where $p(e^2)$, a robust function of the residuals, is defined by

$$p(e^2) = min\left(\frac{e^2}{\sigma^2}, \lambda_3(r-d)\right) \tag{13}$$

where d is the number of dimensions modeled ($d = 3$ for F, $d = 2$ for H), n the total number of features matched across the two frames, k is the number of degrees of freedom ($k = 7$ for F, $k = 8$ for H), r is the dimension of the data ($r = 4$ for 2d correspondences), σ^2 is the assumed variance of the error, $\lambda_1 = log(r)$, $\lambda_2 = log(rn)$, and λ_3 limits the residual error (Torr et al., 1998, Ahmed et al., 2010). Torr uses also the Harris corner detector for feature detection.

4.2.8 Normalized GRIC Difference Criterion The smaller the $GRIC$ score the better the model. If $GRIC_F$ is better than $GRIC_H$, then a good candidate keyframe is indicated. The normalized GRIC Difference Criterion GDC was introduced by Ahmed et. al. (Ahmed et al., 2010) and is defined by:

$$GDC(I_i, I_j)_{x,y} = \frac{GRIC_H(I_i, I_j)_{x,y} - GRIC_F(I_i, I_j)_{x,y}}{GRIC_H(I_i, I_j)_{x,y}} \tag{14}$$

Unfortunately, Ahmed et. al. didn't explain explicit, which kind of feature detection method was used, most likely the Harris corner detector.

4.2.9 Point-to-epipolarline Cost The point-to-epipolarline cost $PELC$ is the standard geometric reconstruction error measure for F given two images I_i and I_j and was named as the Gold-Standard error function by Hartley and Zisserman (Hartley and Zisserman, 2011):

$$PELC(I_i, I_j)_{x,y} = \sum_i d(x_i, \hat{x_i})^2 + d(x_i', \hat{x_i'})^2 \tag{15}$$

This score depends on the chosen feature detection method.

4.2.10 Weighted GRIC and PELC The Weighted GRIC and PELC criterion WGP proposed by Ahmed et. al. represents an alternative keyframe score

$$WGP(I_i, I_j)_{x,y} = w_G GDC(I_i, I_j)_{x,y} + w_P(\sigma - PELC(I_i, I_j)_{x,y}), \quad (16)$$

where σ is the assumed standard deviation of the error. The weights w_G and w_P are not specified by the authors and were set experimentally (Ahmed et al., 2010).

4.3 Shot Boundary Detection

Sometimes uncorrelated frame sequences can be produced while recording videos. This can happen, if the frame rate is very low and a large camera motion becomes somewhat arbitrary, or if the camera has been stopped and then started again at a new position. Shot boundaries are detected by evaluating the correlation between adjacent frames after global motion compensation (Sec. 4.2.6) (Nistér, 2001). The threshold for the Normalized Correlation Constraint is set by the authors to $T_{SB} = 0.75$.

4.4 Videogrammetry in Archaeo3D

In our experience with recording data while moving, videogrammetry is the more fault-tolerant, more cost-effective and easier-to-use approach. The software *JKeyFramer*, an automatic key frame selection tool, was one of the most important outcomes of the project *Archaeocopter*[2]. This tool uses the presented videogrammetric methods for image selection and combines them depending on the objective and was at that time an important step towards fast 3d reconstruction. Meanwhile, it has evolved to allow us to render fast preview models on site. Within the scope of the *Archaeocopter* project, the semi-automatic software *Archaeo3D* was developed to optimize and control the complete reconstruction process. Videos and photos are automatically imported and processed. The software is able to reorder or change the pipeline modules and adjust the parameters, according to the current hardware and the real recording situation and complexity. A combination of *VisualSFM*[3], *COLMAP*, *CMPMVS* and *Meshroom*[4] provided the backbone of the processing toolchain, in all *Archaeocopter* related projects. The *Archaeo3D* reconstruction pipeline includes the following processing steps and software packages:

1. Data recording (*GoPro Hero* vidoes or photo sets)
2. Keyframe extraction (VLC[5], *MPlayer*[6], ffmpeg[7], JKeyframer)
3. Image undistortion (*OpenCV*[8], *JUndistortion*)
4. Image enhancement (*JResizer*, *JEnhancer*)
5. Feature extraction (SiftGPU, *JFeatureManager*)
6. Sparse reconstruction (*VisualSFM*, *COLMAP*)
7. Dense reconstruction (*CMVS+PMVS* (Furukawa and Ponce, 2010), *OpenMVS*[9])
8. Compare or reduce point cloud (*CloudCompare*[10])
9. SGM, Surface fitting (Poisson reconstruction (Kazhdan et al., 2006), *CMPMVS* (Jancosek and Pajdla, 2011), *Meshroom*, *OpenMVS*)
10. Producing orthoimages (*CMPMVS*)
11. Georeferencing, mesh cleaning (*MeshLab* (Cignoni et al., 2008))
12. Integrate data into GIS (*QGIS*[11])

Additional software components like *JUndistortion*, for automatic camera calibration, and *JKeyFramer*, for automatic key frame selection, were developed and integrated. The pipeline automatically shifts processing toward CPU or GPU, depending on the hardware, on which *Archaeo3D* is running. The number of parallel processing jobs is chosen according to the available system memory. While reprocessing old data and preparing new recording campaigns, we also made progress, both in terms of reliability and quality of 3d results, by preparing our software packages *JKeyframer*, *JUndistortion*, *JResizer*, *JFeatureManager* and *JEnhancer*, and releasing them one by one as freely available software tools[12]. The georeferencing step, following the 3d reconstruction process, is an important step due to the fact that 3d models without spatial reference or scale are of limited scientific value. In the *Archaeo3D* workflow, the free software package *QGIS* fulfills this task. As an alternative, the point cloud can also be georeferenced in *VisualSFM*.

Our *Archaeo3D* pipeline allows us to produce preview point clouds and rapidly examine them on-site with the benefit of validating the results immediately. The final reconstruction with *Archaeo3D* off-site, with more powerful computing equipment, will produce more detailed results. This technique was first used during the campaign in Tamtoc/Mexico 2013 (Block et al., 2015). Initially, a number of point clouds of an Huastec settlement site were produced, computed and validated on-site, and afterwards the complete 3d model was produced in the computer lab of the HTW Dresden. We are currently on the way to integrate some parts (such as Keyframe extraction or Image undistortion and enhancement) of this pipeline into our BIM co-registration process.

5. ENHANCED MATCHING ALGORITHM

As it was shown in our previous work the developed co-registration procedure is able to deliver registration accuracies in the range of 3-5 cm. The crucial point of the whole process is the creation of correct line plane pairs. When using manually assigned line plane pairs, it could be shown that even better registration accuracies can be reached. This can be explained with the user's scene understanding. When choosing the segments manually, longer and therefore more stable 3d line segments can be selected. Besides of that, the distribution of selected 3d lines can be more balanced so that ideally line segments are chosen from the entire scene. This also delivers more reliable transformation parameters.

[2] https://www.archaeocopter.de
[3] https://ccwu.me/vsfm/
[4] https://alicevision.org/#meshroom
[5] https://www.videolan.org/vlc/
[6] httsp://www.mplayerhq.hu
[7] https://www.ffmpeg.org/
[8] https://opencv.org
[9] https://cdcseacave.github.io/openMVS/
[10] https://www.danielgm.net/cc/
[11] https://www.qgis.org
[12] https://www.archaeocopter.de

Consequently, it can be said that a reliable classification of the 3d line segments into spatially belonging clusters is of great importance for the automated line plane matching, having the overall aim to get better line plane pairs, in mind. Since SfM reconstructions are not up to scale without further information (e.g. by using control points), the clustering is a challenging task because no metric threshold values can be used. As rotations are scale invariant the direction vectors of 3d line segments play an important role in this context.

The existing solution uses a clustering approach based on established plane hypotheses or rather normal vectors hypotheses. For improving the matching algorithm, we are currently following another approach. Figure 3 shows our test data set covering an indoor scene.

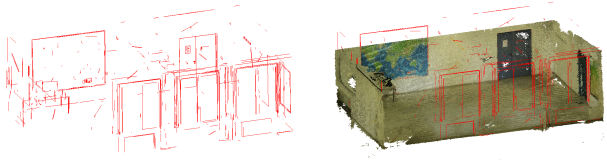

Figure 3. Extracted 3d line segments (left) and dense point cloud for the test room (right).

Since the built environment in large parts follows a Manhattan World (Coughlan and Yuille, 2000), we can calculate the main axes of the reconstructed scene in the point cloud coordinate system by applying a principal component analysis on the direction vectors of the 3d line segments. After finding the main axes, the 3d line segments that are parallel to the main axes are determined using the dot product (see Figure 4).

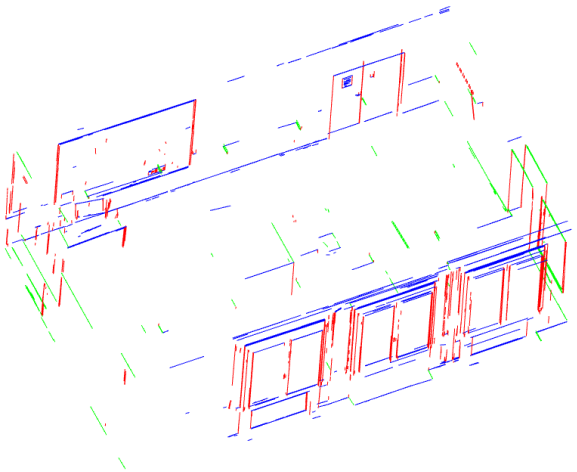

Figure 4. Calculated main axes are displayed in red, green, blue. The parallel lines are colored equal.

In the next step, the distance from the main axes to each mid point for all non-parallel lines are calculated and stored in a list. This list is classified using the *Jenks Natural Breaks* algorithm (Jenks and Caspall, 1971). This clustering algorithm, which is applicable for one dimensional data, tries to group the entries in a way that the variance of the data points inside a group is minimized whereas the variance between the groups is maximized.

An important characteristic of the Jenks algorithm is that it is necessary to specify the number of cluster before running the algorithm. By default, we set the number of clusters to six $nc = 6$. However, using Jenks algorithm it is possible to calculate the *goodness of variance fit* (GVF) ranging from 0 (indicating a bad fit) to 1 (meaning a good fit) which is a quality measure for the evaluation of the clustering. Before that, the *sum of squared deviations for array mean* (SDAM) and the *sum of squared deviations for class mean* (SDCM) need to be calculated for the Jenks clusters:

$$SDAM = \sum_{x \epsilon L} (x - \mu)^2 \tag{17}$$

where L is the list of values to cluster, x represents a single value in L and μ is the mean of L

$$SDCM = \sum_{i=1}^{nc} (x - \mu_i)^2 \tag{18}$$

where nc is the number of clusters, x represents a single value in cluster i and μ_i is the mean of cluster i:

$$GVF = \frac{SDAM - SDCM}{SDAM} \tag{19}$$

Using the quality measure GVF we are increasing the number of clusters as long as $GVF \geq 0.995$. As a result (see 5) we obtain 6 clusters roughly equal to the six main bounding surfaces of the room.

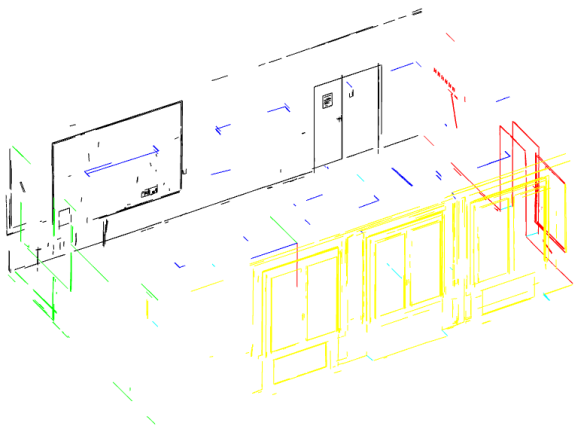

Figure 5. 3d line segments grouped into six major clusters roughly equal to the main bounding surfaces.

After establishing the cluster, the remaining procedure is quite similar to the existing one. We first randomly select three lines from three different clusters. The fourth line is chosen from one randomly chosen cluster that is opposite of one used cluster. So in total we have 4 lines that are matched to all possible sets of four different BIM planes. This process is repeated for a fixed number of times among other things depending on the present room geometry and the resulting minimal configurations are further processed during the adjustment calculation.

6. CONCLUSION AND OUTLOOK

In this article we presented two extensions for the co-registration of image blocks with BIM. For videogrammetric measurements, procedures for optimized image selection were discussed and an overview of the video processing up to the dense point cloud was given. After that, we introduced an improved matching algorithm for the matching of 3d lines (from images) to 3d planes (from BIM). With the new cluster approach, the number of possible matching candidates is reduced. This speeds up the computing time.

The approaches must now be tested further with more complex data. Also, we are currently developing a web service and user interface so that the pipeline can be accessed online.

ACKNOWLEDGEMENTS

This research was funded by the Saxon State Ministry for Science, Culture and Tourism (SMWK), Funding No. 100589640 *Semi-autonomous building inspection with videogrammetry and digital building models (VideoBIM)*. This research is co-financed with tax revenue from the budget approved by the Saxon state parliament.

REFERENCES

Ahmed, M. T., Dailey, M. N., Landabaso, J. L., Herrero, N., 2010. Robust key frame extracton for 3d reconstruction from video streams. *Proceedings of the International Conference on Computer Vision Theory and Applications*, SciTePress - Science and and Technology Publications, 231–236.

Bay, H., Tuytelaars, T., van Gool, L., 2006. Surf: Speeded up robust features. A. Leonardis, H. Bischof, A. Pinz (eds), *Computer Vision – ECCV 2006*, Lecture Notes in Computer Science, 3951, Springer Berlin Heidelberg, Berlin, Heidelberg, 404–417.

Block, M., Ducke, B., Mora Martinez, E., Kroefges, P. C., Rojas, R., Suchowska-Ducke, P., 2015. Low-cost and efficient, uav-based 3d videogrammetry in tamtoc/mexico. *20th European Maya Conference, The Maya in a Digital World (EMC 2015) Bonn, Deutschland*.

Bosché, F., 2012. Plane-based registration of construction laser scans with 3D/4D building models. *Advanced Engineering Informatics*, 26(1), 90–102.

Capel, D., Zisserman, A., 1998. Automated mosaicing with super-resolution zoom. *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, IEEE Comput. Soc, 885–891.

Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G. et al., 2008. Meshlab: an open-source mesh processing tool. *Eurographics Italian chapter conference*, 2008, Salerno, Italy, 129–136.

Coughlan, J. M., Yuille, A. L., 2000. The manhattan world assumption: Regularities in scene statistics which enable bayesian inference. Todd K. Leen, Thomas G. Dietterich, Volker Tresp (eds), *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, MIT Press, 845–851.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381–395.

Furukawa, Y., Ponce, J., 2010. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8), 1362-1376.

Harris, C., Stephens, M., 1988. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*, 147–151.

Hartley, R., Zisserman, A., 2011. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Hofer, M., Maurer, M., Bischof, H., 2017. Efficient 3D scene abstraction using line segments. *Computer Vision and Image Understanding*, 157, 167–178.

Jancosek, M., Pajdla, T., 2011. Removing hallucinations from 3D reconstructions. *Technical Report CMP CTU*.

Jenks, G. F., Caspall, F. C., 1971. Error on Chloroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 61(2), 217–244.

Kaiser, T., Clemen, C., Maas, H.-G., 2022. Automatic co-registration of photogrammetric point clouds with digital building models. *Automation in Construction*, 134, 104098.

Kanatani, K., Ohta, N., 1999. Accuracy bounds and optimal computation of homography for image mosaicing applications. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, IEEE, 73–78 vol.1.

Kazhdan, M., Bolitho, M., Hoppe, H., 2006. Poisson surface reconstruction. *Proceedings of the fourth Eurographics symposium on Geometry processing*, 7.

Kim, C., Son, H., Kim, C., 2013. Fully automated registration of 3D data to a 3D CAD model for project progress monitoring. *Automation in Construction*, 35, 587–594.

Kropp, C., Koch, C., König, M., 2018. Interior construction state recognition with 4D BIM registered image sequences. *Automation in Construction*, 86, 11–32.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), 1615–1630.

Miksik, O., Mikolajczyk, K., 2012. Evaluation of local detectors and descriptors for fast feature matching. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2681–2684.

Morgenthal, G., Rodehorst, V., Hallermann, N., Debus, P., Benz, C., 2021. *Bauwerksprüfung gemäß DIN 1076 – Unterstützung durch (halb-)automatisierte Bildauswertung durch UAV (Unmanned Aerial Vehicles – Unbemannte Fluggeräte)*.

Nistér, D., 2001. Frame decimation for structure and motion. G. Goos, J. Hartmanis, J. van Leeuwen, M. Pollefeys, L. van Gool, A. Zisserman, A. Fitzgibbon (eds), *3D Structure from Images — SMILE 2000*, Lecture Notes in Computer Science, 2018, Springer Berlin Heidelberg, Berlin, Heidelberg, 17–34.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T. et al., 1988. Numerical recipes in C.

Rashidi, A., Dai, F., Brilakis, I., Vela, P., 2013. Optimized selection of key frames for monocular videogrammetric surveying of civil infrastructure. *Advanced Engineering Informatics*, 27(2), 270–282.

Royer, E., Lhuillier, M., Dhome, M., Lavest, J.-M., 2007. Monocular Vision for Mobile Robot Localization and Autonomous Navigation. *International Journal of Computer Vision*, 74(3), 237–260.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-Motion Revisited. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seo, Y.-H., 2008. Optimal keyframe selection algorithm for three-dimensional reconstruction in uncalibrated multiple images. *Optical Engineering*, 47(5), 053201.

Torr, P., Fitzgibbon, A. W., Zisserman, A., 1998. Maintaining multiple motion model hypotheses over many views to recover matching and structure. *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, Narosa Publishing House, 485–491.

Torr, P. H. S., 1998. Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 356(1740), 1321–1340.

Tuttas, S., Braun, A., Borrmann, A., Stilla, U., 2017. Acquisition and Consecutive Registration of Photogrammetric Point Clouds for Construction Progress Monitoring Using a 4D BIM. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 85(1), 3–15.

Vincke, S., Vergauwen, M., 2020. Geo-Registering Consecutive Construction Site Recordings Using a Pre-Registered Reference Module. *Remote Sensing*, 12(12), 1928.

Wu, C., 2010. SiftGPU: A GPU Implementation of Scale Invariant Feature Transform (SIFT). http://www.cs.unc.edu/ ccwu/siftgpu/.

Wujanz, D., Schaller, S., Gielsdorf, F., Gründig, L., 2018. Plane-Based Registration of Several Thousand Laser Scans on Standard Hardware. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2, 1207–1212.