IMAGE-BASED CONTROL POINT DETECTION AND CORRESPONDENCE-FREE GEOREFERENCING

Christian Benz[a,*], Volker Rodehorst[a]

[a]Computer Vision in Engineering, Bauhaus-Universität Weimar, Germany – christian.benz@uni-weimar.de

KEY WORDS: Control Point Detection, Semantic Segmentation, Deep Learning, Georeferencing, Point Cloud Registration.

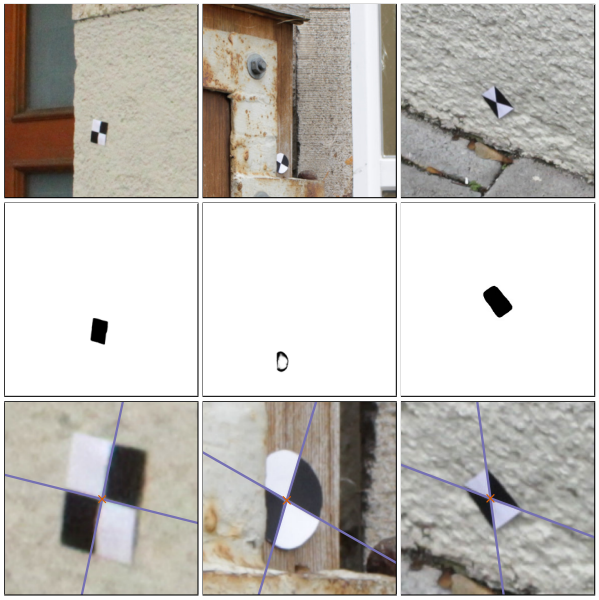
ABSTRACT:

In order to appropriately measure properties in 3D models, accurate georeferencing plays a vital role in structural health monitoring. For that purpose, control points are attached to the surface of the structure and measured geodetically. These points can be recovered in the virtual model and associated with the geodetic measurements. Automating the process of detecting and associating control points and geodetic measurements, facilitates the accurate georeferencing of large 3D models. While the number of marker types for control points is steadily increasing, this work claims that – under a plausible assumption – comparatively simple and commonly used marker designs can serve for accurate and robust georeferencing. By assuming that control points are asymmetrically distributed over the surface of the structure, the correspondence of points is determined by their geometric interrelation. In this work, an image-based detector for relatively simple control point types is proposed, applying transfer learning on hierarchical multi-scale attention (HMA) (Tao et al., 2020). For associating detected and geodetically measured points, a RANSAC-based procedure is presented that determines a geometrically consistent transformation between detected and measured points.

1. INTRODUCTION

The automation and digitization of processes in the field of structural health monitoring is picking up pace. Images are a key ingredient to this process, since they form the basis for 3D reconstruction and the maintenance of an information-rich and vivid digital representation of the structure. Moreover, the field of image-based object recognition has made noticeable progress during the last decade. This is partly due to the availability of larger datasets and the (re-)introduction of artificial neural networks. Making reasonable use of these techniques for purposes of structural health monitoring can benefit all parties involved.
Performing measurements in the virtual model is essential for assessing health-critical properties, such as the level of structural deformation or the size of defects. The unit of measurement is, however, only meaningful, if the model is properly georeferenced. A prerequisite for accurate referencing is precise knowledge of the metric relations within the structure. This knowledge is, typically, unavailable apriori. On that account, salient points (control points, targets, fiducial markers) are attached to the surface of the structure. On site, these points are measured by geodetic experts using professional equipment, in order to locate the control points in a geodetic reference frame.

Incorporating the measured control points into the virtual model usually requires manual interaction. Some providers of 3D reconstruction software offer specifically encoded marker types for printout, that – attached to the structure – can be automatically detected by the software itself. This, however, restricts the usage to a specifically customized marker type. Being tied to one specific software for marker detection, also noticeably limits the freedom of the user to choose a preferred software for reconstruction and georeferencing. Thus, customized marker types might not render a general solution.
The number of marker designs proposed by the community constantly increases, including novel types of ID encodings. While



(a) (b) (c)

Figure 1. Qualitative results of control point detection. Top row: input images, central row: prediction, bottom row: center refinement.

for many applications ID encoding markers are of high value, it is here demonstrated that georeferencing of a structure can be accomplished without complex marker designs. Instead of increasing the complexity of the marker pattern, complexity is added to the process of relating marker occurrences in the virtual model and on-site measurements. By shifting complexity to later stages, a more general procedure of automated georeferencing is achieved.

Based on this contextual frame, the contributions of this work are threefold: (1) Transfer learning on the state-of-the-art ap-

* Corresponding author.

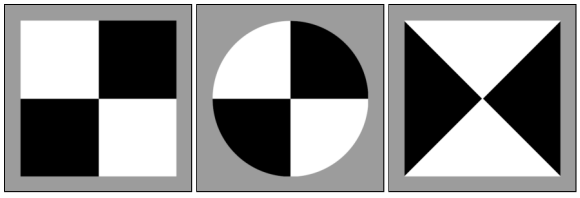


Figure 2. Considered types of control points.

proach of hierarchical multi-scale attention (HMA) (Tao et al., 2020) is demonstrated to achieve good results for simple marker types. (2) A procedure is presented, that involves Hough transform for subpixel accurate localization of the control point center. (3) In order to compensate for lack of ID encoding in the markers, RANSAC is adjusted to establish robust correspondences between (virtually) detected and (actually) measured control points. RANSAC, thereby, yields a plausible transformation, that can serve as reasonable initial guess for further registration procedures.

2. RELATED WORK

The related work covers current approaches to control point detection, convolutional neural networks (CNN), and techniques for point cloud registration.

2.1 Control Point Detection

The detection of control points (CP) or fiducial markers is relevant in a number of fields, i.a. in virtual reality, 3D reconstruction, drone navigation, or georeferencing. For the growing field of drone technologies, some providers of related software offer a module for automated control point detection. PIX4D provides the module *AutoGCP* to overcome the manual annotation of ground control points (GCP) in the acquired imagery[1]. Including the checkerboard-based GCP pattern, they cover so-called *AeroPoints*, which are portable ground stations with GNSS antennas to obtain accurate global positioning[2]. METASHAPE has added a type of coded target marker, which resembles the ShotCode approach, cf. (Kato and Tan, 2007). Before capturing a site photographically, the markers need to be attached in the region of interest. Following a unique coding schema, each target has an individual circular coding, which in case of sufficient visibility is subject to automated image-based detection and unique marker identification. DRONEDEPLOY offers a Ground Control AI feature to automated control point detection[3]. By using the annotations provided through user corrections of wrong predictions, they attempt to build up a real-world training base. Besides checkerboard-based control points, other types are covered such as cross-based and color-related patterns.

Apart from the commercial approaches, a handful of coding projects are publicly accessible. Most notably GEOBITS[4] provide an approach to automated ArUco marker detection, that is based on the OpenCV ArUco library (Garrido-Jurado et al., 2014). The library offers basic and advanced functionality on the creation and detection of ArUco markers and pose inference. In science, (Rumpler et al., 2014) presented a workflow of precise and geo-accurate reconstruction involving the design of ID encoding fiducial markers. Furthermore, the approach is based on the assumption of known correspondences between detected and geodetically measured markers. (Yu et al., 2020) provide an overview over the various different marker designs that have evolved over the last decades. Moreover, a detection algorithm for the invented TopoTag is presented, that applies i.a. thresholding, binarization, and topological filtering to obtain a robust detection. (DeGol et al., 2017) indicate that approaches not based on machine learning, might be faster due to their potential independence from GPU computing resources.
The presented current developments in the field of control point and fiducial marker detection regularly involve the design of novel marker types and corresponding detection algorithms. The assumption in the here presented work is, that simpler control point types are sufficient for automated georeferencing when exploiting the disambiguating power of unique intrinsic geometric relation between the point clouds. Simpler marker types do, furthermore, facilitate the establishment of a convention in control point usage at structures.

2.2 Convolutional Neural Networks

Starting with the victory in the ImageNet challenge in 2012 (Deng et al., 2009, Krizhevsky et al., 2012), the application and research of convolutional neural networks (CNN) has constantly increased in science as well as in industry. Leading approaches in visual recognition tasks, such as image classification, object detection, or semantic segmentation are based on CNNs, e.g. (Cordts et al., 2016, Lin et al., 2014).
Convolutional neural networks are a specific kind of artificial neural networks that make extensive use of the convolution operation. The convolution operation, due to its spatial properties, has proven especially suitable for image applications. In a cascade of convolutional and pooling layers the CNN maps the input image to an output. In the case of image classification, the output represents the candidate classes for the input image. For semantic segmentation the output is of same height and width as the input and contains pixelwise, dense class predictions. (Long et al., 2015) demonstrated that semantic segmentation can be comparatively naturally approached with fully convolution networks. Fully convolutional networks do not contain fully connected layers. Thereby, they become relatively independent of specific input sizes and can potentially handle images of various height and width.
Since artificial neural networks are trained in supervised fashion they require a representative and sufficiently large dataset. For many domains, however, no such datasets are available. In order to overcome the lack of data, a number of methods have evolved, including data augmentation (Shorten and Khoshgoftaar, 2019, Dwibedi et al., 2017), transfer learning (Tan et al., 2018), and the incorporation of synthetic data (Richter et al., 2016). Benchmarking challenges such as Cityscapes (Cordts et al., 2016), COCO (Lin et al., 2014), or ADE20k (Zhou et al., 2019) form a solid basis to find proper approaches for transfer learning.

2.3 Point Cloud Registration

The procedure most commonly referred to in point cloud alignment is *iterative closest point* (ICP) (Besl and McKay, 1992, Pomerleau et al., 2015). For each step of the iteration and each point in the source cloud, the closest point in the target cloud is computed. The corresponding point pairs between source and target point cloud serve as basis for estimating the transformation between the source and target. After a sufficient number

[1] https://www.pix4d.com/blog/automatic-ground-control-points
[2] https://www.propelleraero.com/aeropoints/
[3] https://support.dronedeploy.com/docs/automated-gcp-detection-with-dronedeploy
[4] https://github.com/dronemapper-io/aruco-geobits

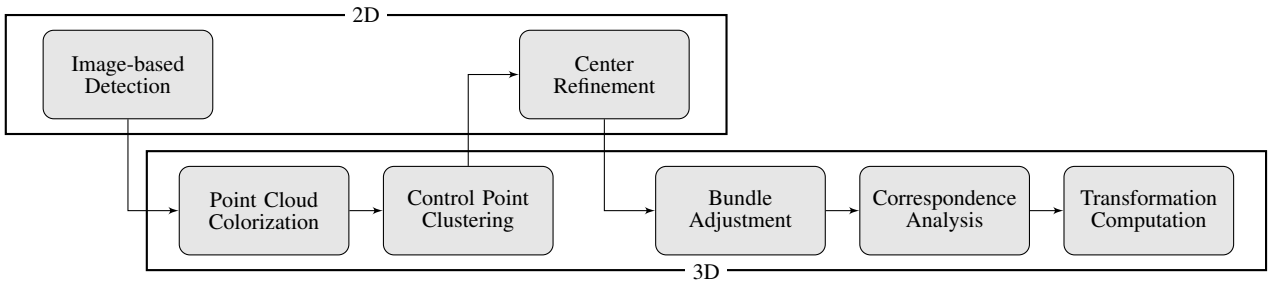


Figure 3. Schematic workflow of control point detection, refinement, correspondence analysis, and transformation computation.

of iterations, the algorithm returns an accurate transformation between the point clouds. ICP is based on the assumption of a proper initial alignment. Lacking this initialization, the registration can run into undesirable transformations based on inappropriate correspondences. Furthermore, the original ICP is restricted to rigid transformations, since e.g. variance in scale would impede with the closest point computation.
To overcome the initialization constraint (Gelfand et al., 2005) estimate the initialization based on local geometry descriptors. Subsequently ICP is executed. In order to obtain a global registration, (Aiger et al., 2008) and (Makadia et al., 2006) use approaches based on RANSAC (Fischler and Bolles, 1981): Point correspondences between source and target cloud are randomly selected, there based upon, a transformation is estimated, and, eventually, the difference between transformed source cloud and target cloud is measured. This procedure is repeated until a well-aligning transformation has been found.

3. DATA

Data form an essential ingredient to data-driven approaches such as deep learning. The control point types, the image data, and the 3D object for testing are presented.

3.1 Control Point Types

The number of different control point or fiducial marker types is constantly growing, cf. (DeGol et al., 2017). Different marker properties serve different purposes, e.g. color for fast misdetection rejection, ID encoding for unique identification, sharp crossings for precise aiming. For many application fields, the variety of types impede the establishment of a convention. For structure inspection, however, it is here claimed that the comparatively simple checkerboard-based marker pattern can suffice, cf. Figure 2. The markers show a simple, intuitive pattern and are presumedly cheap in print and attachment. They have a clear, well-defined aiming point, and they are the currently most prevalent ones. They can, optionally, obtain a character-based identifier in the margin, which, however, is not necessary for the purpose of georeferencing. Figure 2 depicts the checkerboard-based CPs considered: a square 2×2 pattern, a circular 2×2 pattern (Secchi disk), and a squared 2×2 version with 45° pattern rotation.

		Training	Validation	Testing
S2DS	Marker class	70	9	10
	Other classes	493	78	83
Synthetic	Various bridges	–	–	546
Various	Out of dataset	–	–	38
Total		563	87	677

Table 1. Images used for training, validation, and testing.

3.2 Dataset

In order to cover a variety of backgrounds and acquisition setups, three sources of image data are used for training and evaluation, cf. Table 1. Training is based on the *structural defect dataset* (S2DS), which consists of a total of 743 images and corresponding segmentation annotations. It represents seven classes, including five types of structural defects (crack, spalling, corrosion, efflorescence, vegetation), background, and control point. All images were acquired at real inspection sites and are cropped to size 1024×1024. The dataset contains a large number of images without markers, which supports modeling the negative class and reduces the bias on the prior. The images differ with respect to image quality (sharpness, color constancy, lighting condition, etc.) and acquisition devices (mobile phone, DSLR/DSLM, drone).
Moreover, 546 images were synthesized and used for testing. In order to reduce the gap to real images, control points of the above presented types were rendered onto images from actual bridges. The control points are furthermore morphed and assimilated to the background and endowed with noise and blur. The synthetic nature of the images is still perceptible to the human eye.
For assessment of the generalization capabilities of the approach, a dataset with distinctly different backgrounds compared to S2DS was created. While the majority of images from S2DS shows surfaces of concrete structures, the *various* dataset contains images from gray and red stone, plaster and roughcast. Furthermore, it shows the different marker types in various sizes and from various angles.

3.3 Bridge Pier

The entire workflow was carried out on various examples and, due to space constraints in this paper, demonstrated using one exemplary 3D reconstruction of a real bridge pier. The used segment is approximately 10 meters high and was reconstructed with 1460 images using the commercial Metashape software. The images were captured by a DJI drone carrying a Zenmuse P1 full-frame camera with focal length 35mm. Each image is roughly 45MP large (8192×5460 px). For the relevant part of the bridge, the acquisition distance alternated around 2 meters. The model contains nine of the above specified marker types, which are asymmetrically distributed over the surface.

4. DETECTION

The detection of control points and their accurate 3D localization involves a number of processing steps. Figure 3 provides an overview of the proposed workflow. After image-based detection, the results are transferred into the 3D space using dense point cloud colorization. Based on the centroids from clustering, the centers of the control points can be refined in the 2D domain. The following stages are conceptually closer to georeferencing and, thus, referred to in Section 5.

4.1 Image-based Detection

In order to achieve a robust performance for control point detection given the not abundant amount of data, the *hierarchical multi-scale attention* (HMA) approach by (Tao et al., 2020) is used for transfer learning. At the time of experimentation, HMA was the highest ranked approach in the pixel-level semantic labeling task of the Cityscapes benchmark (Cordts et al., 2016) with accessible code. It is currently occupying the 8$^{\text{th}}$ position in the challenge and still the second highest ranked approach with published code. The highest ranked approach with code is an extension of HMA with boundary-aware loss (Borse et al., 2021), which outperforms HMA by a small margin.
Figure 4 schematically depicts the training and inference procedure of HMA. During training, the image is fed to the network in two scales. Both scales pass through the backbone, the HRNet-OCR (Yuan et al., 2020, Sun et al., 2019), which refers to the high-resolution network with object-contextual representations. The semantic head consists of the following layers (3×3 conv) → (BN) → (ReLU) → (3×3 conv) → (BN) → (ReLU) → (1 × 1 conv), where conv refers to the convolutional layer, BN to the batch norm, and ReLU to the rectified linear unit. The attention head is structurally similar to the semantic head and yields a probability map. This probability map determines the attention, that is supposed to be paid to the scale with respect to the other scale. Since attention is contrastively learned for two scales, HMA generalizes to an arbitrary number of scales in inference: The attention is applied to a scale pair, while additional scales are recursively incorporated into the accumulation process.
For training, the region mutual information (RMI) loss is used (Zhao et al., 2019), which consists of a cross-entropy and mutual information component. While cross-entropy assesses each pixel independently, mutual information considers the local neighborhood around the pixel. It, thereby, accounts for slight misalignments in the prediction. Extensive data augmentation is used, including scaling, rotation, shifting, and blur for 80% of the samples.

4.2 Cloud Colorization and Clustering

In order to transfer the 2D information from image-based detection into the 3D space of the model, cloud colorization is performed. For that purpose, each point of the dense point cloud is projected into all images of the scene. The point is typically invisible in many images, while some images – due to proximity or quality – provide particularly reliable information about the point. The information is gathered from all images and fused into a single response for each point, whether the point represents or not represents a control point.
It is assumed, that points in a local neighborhood are regularly assigned to the same class. That is, points of the cloud, that represent control points, are supposed to cluster locally. In order to extract clusters, the DBSCAN algorithm ("density-based spatial clustering of applications with noise" (Ester et al., 1996) is applied: Based on a spatial density criterion, samples are tagged as either in- or outliers. All inliers of a cluster are presumed to represent a control point. The centroid fo the cluster serves as preliminary 3D location of the control point.

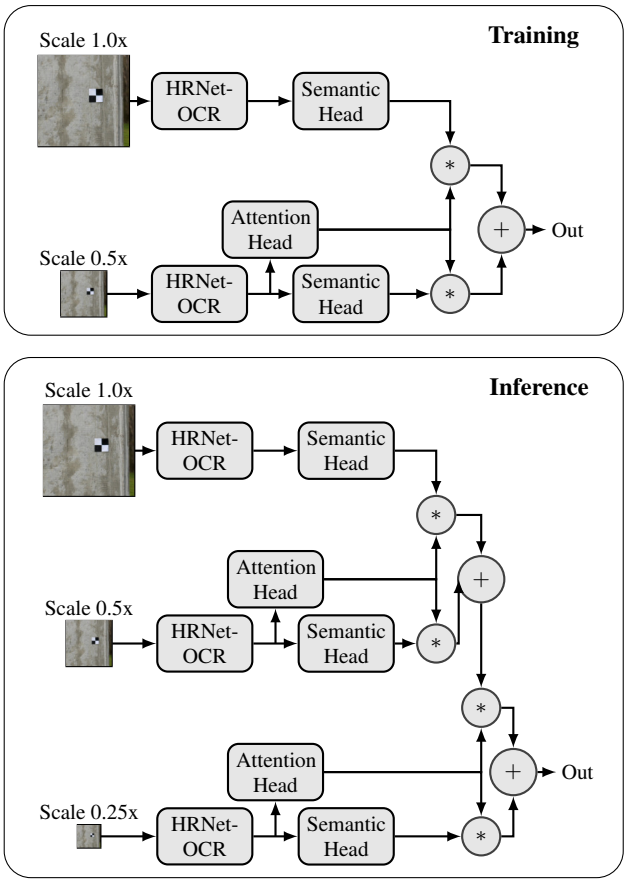


Figure 4. Depiction of the training and inference procedure of hierarchical multi-scale attention (HMA).

4.3 Center Refinement

The result from centroid computation after clustering serves as a rough estimate of the 3D position of the control point. In order to refine the localization, a dedicated search for the marker's center is performed on all images, in which the control point occurs. For that purpose, the centroid is projected into all images and Hough transform (Hough, 1962, Duda and Hart, 1972) for detecting straight lines is computed on the patch around the projected point. Determining the intersection point of roughly orthogonal lines, yields a subpixel accurate estimate of the center.
Equation 1 denotes the normal representation of a line as assumed for Hough transform. Angle θ and distance ρ span Hough transform's two-dimensional parameter space. Straight lines are determined based on a voting scheme, in which more votes correspond to higher evidence for the occurrence of such a line.

$$x\cos(\theta) + y\sin(\theta) - \rho = 0 \tag{1}$$

$$l_1 = [\cos(\theta_1), \sin(\theta_1), -\rho_1]^T \tag{2}$$

$$l_2 = [\cos(\theta_2), \sin(\theta_2), -\rho_2]^T \tag{3}$$

$$45^\circ < |\theta_1 - \theta_2| \bmod 180^\circ \leq 135^\circ \tag{4}$$

$$l_1 \times l_2 = p \tag{5}$$

The straight line with the highest number of votes forms l_1 in the homogeneous line representation. The second line must fulfill the relaxed orthogonality constraint, Equation 4, i.e. requires orientation between 45° and 135° with respect to l_1. The intersection point p results from computing the cross product of l_1 and l_2.

5. GEOREFERENCING

Beside the automated detection of control points, the automated association of these detected control points with the geodetically measured ones, is a key aspect of automated georeferencing. Figure 3 shows the workflow of detection and georeferencing. In bundle adjustment, the deviations of multiple occurrences of a control point are balanced over points and projection matrices. The RANSAC-based correspondence analysis is followed by the final transformation computation.

5.1 Ray Casting and Bundle Adjustment

Center refinement yields the precise localization of the targeting point on image level. In order to determine the corresponding 3D position *ray casting* (Roth, 1982) is performed. In ray casting a virtual ray is casted from the camera origin through the respective pixel into the 3D scenery. The point of intersection with the 3D model serves as 3D coordinate. Due to image overlap, which is fundamental to photogrammetric 3D reconstruction, one control point can occur in multiple images. Applying ray casting leads to multiple coordinates in 3D space for one control point. Depending on the accuracy of the reconstructed model and camera poses, these control point centers are typically located in close proximity. Which 3D occurrences represent the same control point can be inferred from control point clustering earlier in the workflow.
In order to settle the error of the various 3D occurrences of one control point, bundle adjustment is applied. In a network of spatial rays spanning from the projection centers through the image points, the global error is subject to minimization:

$$\arg\min_{\mathbf{P}_i,\mathbf{X}_j} \sum_{i=1}^{n}\sum_{j=1}^{m} ||\mathbf{P}_i\mathbf{X}_j - \mathbf{x}_{i,j}||_2 \tag{6}$$

$\mathbf{P}_i$ refers to projection matrix i (out of n), $\mathbf{X}_j$ to 3D object point j (out of m), $\mathbf{x}_{i,j}$ to the measured image point corresponding to $\mathbf{P}_i\mathbf{X}_j$, and $||\dots||_2$ denotes the Euclidean norm. In bundle adjustment both, projection matrices and object points, are optimized, leading to a single occurrence of an object point, i.e. one 3D representation for each control point.

5.2 Correspondence Analysis and Registration

Based on the lack of reasonable initialization of correspondences and transformation, a RANSAC-based approach for registration is applied here. Assuming $\mathbf{S}$ to be the source point cloud and $\mathbf{R}$ to be the reference point cloud, to which $\mathbf{S}$ is supposed to be transformed. Furthermore, be $\mathbf{T}$ a similarity transformation from $\mathbf{S}$ point cloud space into $\mathbf{R}$ point cloud space. Assuming homogeneous coordinates and correspondence-preserving sorting $\mathbf{S}, \mathbf{R} \in \mathbb{R}^{4\times k}$, where k refers to the number of control points, and $\mathbf{T} \in \mathbb{R}^{4\times 4}$. Parameter θ represents the correspondence mapping between $\mathbf{S}$ and $\mathbf{R}$, i.e. relates which points in the cloud are mutually assigned. The optimization objective to solve is:

$$\arg\min_{\theta} \|\mathbf{R} - \mathbf{T}_\theta\mathbf{S}\|_2 \tag{7}$$

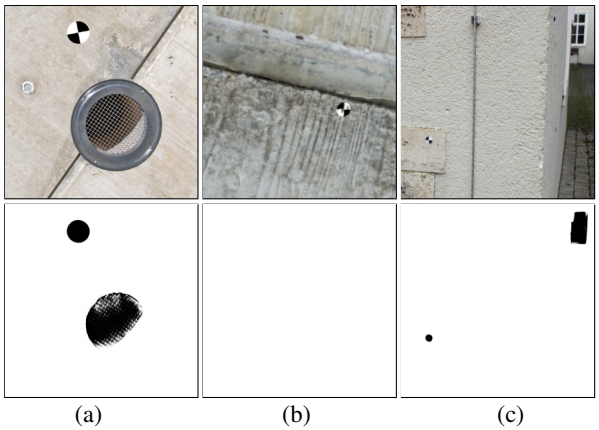
That is, a correspondence mapping θ is to be found that yields a transformation $\mathbf{T}$, which minimizes the Euclidean distances between corresponding points in source and reference point cloud. Assuming a similarity transformation for $\mathbf{T}$, seven parameters for translation, rotation, and scale can be estimated from three point correspondences.
Being a global registration problem, the correspondences in this context are, however, not known. To overcome the lack of correspondence knowledge, RANSAC is applied, i.e. for each iteration a minimal number of correspondences are randomly selected, the transformation is estimated based on these correspondences, and the optimization objective is evaluated based on the inferred transformation. As termination criterion serves the average distance between the point cloud being below 10cm.

6. EVALUATION

In this section a description of the qualitative and quantitative results in the image space is provided. Moreover, the procedure is illustrated in the 3D space of the bridge pier model.

6.1 Detection

Figure 1 qualitatively illustrates the performance of the detector on three example images. The top row shows the input image, the central row the predicted segmentation masks (bright represents low and dark high probabilities), and the bottom row shows the results of the center point location. All three images form representative samples from the challenging *various* dataset. Even though they are captured in non-orthogonal view, both fully visible markers (a and c) are robustly detected. The visible boundaries of the third marker (b) – that is subject to occlusion – are detected, while the center forms false negatives. The center refinement is capable of compensating the misclassification in (b) and yields convincing results for all three images. Failure cases are shown in Figure 5. The two left images are synthesized images, while the right is part of the *various* dataset. For images (a) and (c) false positive detections occur. In (a), a louver is falsely classified as control point. Presumably, the round shape and the dark shades contribute to the classifier's



(a) (b) (c)

Figure 5. Failure cases of control point detection. Top row: input images, bottom row: prediction.

	IoU [%]	F_1 [%]	AP [%]
S2DS	98.2	99.1	99.4
Synthetic	95.1	97.5	99.0
Various	73.7	84.8	78.2
Total	89.0	93.8	92.2

Table 2. Performance of the detection approach on the test data.

misjudgment. The checkerboard corner at the center of the control point, which the human eye might consider salient, seems, interestingly, less import to the classifier. This is also indicated by result (b) from Figure 1, where the center receives false negative detections. In Figure 5 (c) a misclassification occurs for the window in the background with characteristic white frames. The false negative for synthetic image (b) might be due to the texture of the surface that shines through the morphed control point.

Quantitative results are provided in Table 2. The quantitative evaluation took place on the respective test sets. As performance metrics, intersection-over-union (IoU, aka Jaccard score), F_1 score, and average precision (AP) are used. The performance on S2DS and the synthetic dataset is near perfect. Typical misclassification on the synthetic set are the ones referred to above: louver-like artifacts on the image and rough surface texture. The performance on the challenging *various* dataset is distinctly lower than on the other two datasets. That is due to the different backgrounds not represented in the training set, such as stone walls and plaster. Furthermore, as indicated in Figure 5 (c), objects such as remote windows do occasionally occur on the images and lead to confusion.

6.2 Registration

Figure 6 shows the 3D model described in Section 3.3. The blue flags indicate the resulting positions from semantic segmentation, center refinement, ray casting, and clustering. The results were converted into the marker format of and imported and visualized in Metashape. The positions are the ones automatically determined by the mentioned processing steps. The tags accompanying each flag, refer to the ID of the geodetically measured control points. These IDs were unknown and needed to be assigned using the RANSAC-based correspondence analysis.

For the given example the functionality of the procedure is confirmed. Potential obstacles that can occur in larger and less controlled environments predominantly involve false positive and false negative detections. False positives, i.e. detection of a control point where no one was measured, can be caused by confusing patterns to the detector. Confusion may originate in unknown backgrounds and environments or the presence of other high-contrast control points or patterns. Furthermore, control points might have been overlooked during geodetic measurement. False negatives on the other hand refer to the missed detection of measured control points. Reasons for false negatives are low image quality, low marker size in the image, missing images, or occlusions. Redundancy resulting from overlapping images can potentially compensate up to a certain degree for occasional false negatives. Many of mentioned deficits can be resolved beforehand through careful and meticulous image acquisition.

Figure 6. 3D model of a segment of bridge pier with assigned control points.

7. CONCLUSION

In this work a procedure for images-based control point detection and automated georeferencing was proposed. Its functioning was demonstrated on a 3D model of a corner of a bridge pier. Compared to other approaches, the here presented approach is based on relatively few assumptions and restrictions, which alleviates its broader application in science and industry. Unlike other approaches, the one presented here does not require and propose an elaborated, ID encoded type of control point. Rather it builds upon the simple and already widely used checkerboard design, which is intuitive to understand, precisely to target, easy to print and obtain. These properties make it a candidate choice for a control point convention, i.e. the standard control point attached to structures. Even though no ID needs to be encoded into the pattern, IDs at the border of the marker are still valid and appreciable e.g. for the geodetic reference measurements.

Beyond that, the approach does not assume knowledge about correspondences between the virtual and measured control points. It, thus, is *correspondence-free* in that sense. Correspondences are indispensable for georeferencing. They are,

however, intrinsically determined in the process through the proposed RANSAC-based procedure.
Making relatively few assumptions, one assumption, however, needs to be enforced in order for the procedure to work: the *asymmetry* assumption. The control points must be attached in a way to incorporate at least one aspect of asymmetry. Imagine the case where eight control points are perfectly precisely attached to the four corners and the bottom of the legs of an ordinary table. Without knowing at least one correspondence, point cloud registration is not unambiguously achievable. Keeping the asymmetry assumption in mind, there are no major obstacles for the practical application of the proposed approach.
As mentioned, there are libraries available e.g. for ArUco marker detection. This library applies traditional image processing techniques for accomplishing the task. It can be assumed that for the task of detecting the checkerboard-based markers used in this work, traditional approaches do a solid job. Very probably, traditional methods come with the benefit of requiring less computation resources than data-driven approaches. Moreover, the marker type used here is well-defined and effective features are more or less immediately clear for the expert in the field. Thus, a top-down design might also be the theoretically more elegant solution to control point detection.
The approach used for transfer learning, hierarchical multi-scale attention (HMA), is likely oversized for the comparatively easy task of control point detection. HMA, however, occurs appropriate, when control point detection forms only a minor component in detection. The ultimate goal is to learn a multi-class model, that masters not only control point detection, but the detection of various defects. Providing enough capacity for additional object classes HMA renders a suitable approach for inclusion of multiple other classes. That control points can be represented by a rather simplistic type of markers was demonstrated in this work.

ACKNOWLEDGEMENTS

The research in this paper was funded within the AISTEC research project by the German Federal Ministry of Education and Research (BMBF) under the grant number 13N14657.

REFERENCES

Aiger, D., Mitra, N. J., Cohen-Or, D., 2008. 4-points congruent sets for robust pairwise surface registration. *ACM SIGGRAPH 2008 papers*, 1–10.

Besl, P. J., McKay, N. D., 1992. Method for registration of 3-d shapes. *Sensor Fusion IV: Control Paradigms and Data Structures*, 1611, International Society for Optics and Photonics, 586–606.

Borse, S., Wang, Y., Zhang, Y., Porikli, F., 2021. Inverseform: A loss function for structured boundary-aware segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5901–5911.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

DeGol, J., Bretl, T., Hoiem, D., 2017. Chromatag: A colored marker and fast detection algorithm. *Proceedings of the IEEE International Conference on Computer Vision*, 1472–1481.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 248–255.

Duda, R. O., Hart, P. E., 1972. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1), 11–15.

Dwibedi, D., Misra, I., Hebert, M., 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. *Proceedings of the IEEE International Conference on Computer Vision*, 1301–1310.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *kdd*, Vol. 96, No. 34, 226–231.

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.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., Marín-Jiménez, M. J., 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6), 2280–2292.

Gelfand, N., Mitra, N. J., Guibas, L. J., Pottmann, H., 2005. Robust global registration. *Symposium on Geometry Processing*, Vol. 2, No. 3, Vienna, Austria, 5.

Hough, P. V., 1962. Method and means for recognizing complex patterns. US Patent 3,069,654.

Kato, H., Tan, K. T., 2007. Pervasive 2D barcodes for camera phone applications. *IEEE Pervasive Computing*, 6(4), 76–85.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, Springer, 740–755.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Makadia, A., Patterson, A., Daniilidis, K., 2006. Fully automatic registration of 3d point clouds. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1, IEEE, 1297–1304.

Pomerleau, F., Colas, F., Siegwart, R., 2015. A review of point cloud registration algorithms for mobile robotics. *Foundations and Trends in Robotics*, 4(1), 1–104.

Richter, S. R., Vineet, V., Roth, S., Koltun, V., 2016. Playing for data: Ground truth from computer games. *European Conference on Computer Vision*, Springer, 102–118.

Roth, S. D., 1982. Ray casting for modeling solids. *Computer Graphics and Image Processing*, 18(2), 109–144.

Rumpler, M., Daftry, S., Tscharf, A., Prettenthaler, R., Hoppe, C., Mayer, G., Bischof, H., 2014. Automated end-to-end workflow for precise and geo-accurate reconstructions using fiducial markers. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3, 135–142.

Shorten, C., Khoshgoftaar, T. M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1–48.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J., 2019. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. *International Conference on Artificial Neural Networks*, Springer, 270–279.

Tao, A., Sapra, K., Catanzaro, B., 2020. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*.

Yu, G., Hu, Y., Dai, J., 2020. TopoTag: A Robust and Scalable Topological Fiducial Marker System. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*.

Yuan, Y., Chen, X., Wang, J., 2020. Object-contextual representations for semantic segmentation. *ECCV 2020: 16th European Conference on Computer Vision, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, Springer, 173–190.

Zhao, S., Wang, Y., Yang, Z., Cai, D., 2019. Region mutual information loss for semantic segmentation. *arXiv preprint arXiv:1910.12037*.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A., 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3), 302–321.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-399-2022 | © Author(s) 2022. CC BY 4.0 License.