Active vision for 3D indoor scene reconstruction using a 3D camera on a pan-tilt mechanism

Mateja Hržica[a]*, Robert Cupec[a], and Ivan Petrović[b]

[a]*Department of Computer Engineering and Automation, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Kneza Trpimira 2b, Osijek, Croatia;* [b]*Department of Control and Computer Engineering, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia*

*mateja.hrzica@ferit.hr

Active vision for 3D indoor scene reconstruction using a 3D camera on a pan-tilt mechanism

We present a novel approach to automatic indoor scene reconstruction from RGB-D images acquired from a single viewpoint using active vision. The proposed method is designed to select the next view with sufficient information for reliable registration. The next view is selected based on the percentage of unexplored scene regions captured inside the field of view and the information content in the overlapping region between the image acquired from the next view and one of the previously acquired images. It is required that this overlapping region contains surfaces with different orientations, whose alignment provides a reliable estimation of the relative camera orientation. The point correspondences between views are established using the assumption of fixed viewpoint, imprecise information about relative view orientation and local surface normal, without need for features based on texture or distinctive local shape. After completing a scan, the 3D scene model is constructed by performing registration of the acquired depth images. Two algorithms are considered for that purpose: a point-to-plane ICP with point weighting based on the properties of the measurement noise and TEASER++. The proposed method is tested on the synthetic dataset Replica.

Keywords: active vision; next-best-view; iterative closest point registration; point cloud; indoor scene reconstruction

1. Introduction

This paper addresses the problem of obtaining 360° scan of the robot's local environment using a 3D camera mounted on a pan-tilt mechanism (PTM). A 3D model of the local environment in the current robot position can be obtained using various 3D sensors such as 3D LiDAR or a 3D camera. The advantage of the 3D LiDAR over 3D cameras is 360° view angle. On the other hand, 3D LiDARs are much more expensive. Since a 3D camera has a limited field of view (FoV) and thus can capture only a region in the robot's environment, a fusion of multiple depth images is necessary to obtain a

complete scene model. The main problem in fusion of multiple images is their registration, i.e. determining the camera poses from which the images are acquired with respect to a common reference frame (RF).

In this paper, we propose a novel active vision strategy for the automatic creation of a local 3D scene model from a particular viewpoint using a 3D camera, where the camera FoV is expanded using a consumer-grade pan-tilt mechanism (PTM). It is assumed that the pan and tilt angle of the camera are known, but the uncertainty of this information is too high to allow building of accurate scene models without image registration. The proposed method is designed to select the next view with sufficient information for reliable registration of the new view with the previously acquired depth images. At the same time, the new FoV must contain as many unexplored parts of the scene as possible.

The proposed active vision algorithm is integrated into a 3D scene reconstruction system, specially designed for indoor environments characterized by uniformly coloured featureless flat surfaces, such as walls, floor, ceiling, doors, tables etc., where standard keypoint detectors cannot be relied on. Instead, we use inaccurate camera pose information provided by the PTM, local surface normal orientation and fixed viewpoint constraint to establish correspondences between points lying on featureless flat surfaces. The experiments presented in this paper demonstrate that even in this conditions accurate 3D models can be obtained.

Although the next-best-view (NBV) selection approach is the main contribution of this paper, in order to demonstrate its usefulness, we present a complete 3D modelling framework, which uses the proposed approach to acquire depth images and produces accurate environment models as the final result. A complete system overview of the proposed 3D indoor scene reconstruction system is shown in Figure 1.

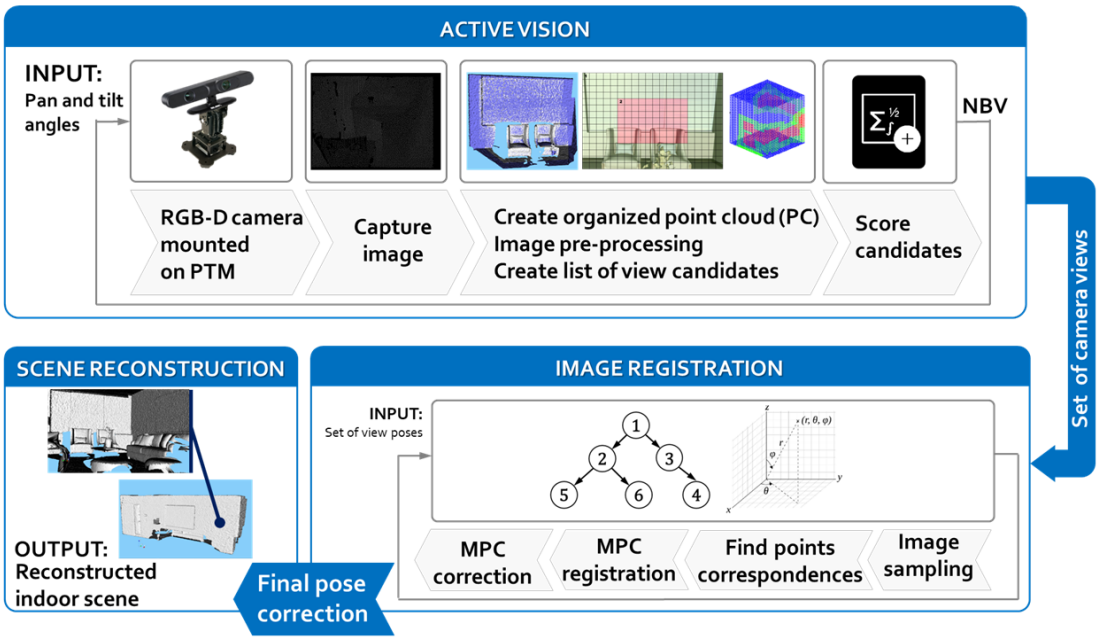


Figure 1. Overview of the proposed method for 3D indoor scene reconstruction using a 3D camera on a pan-tilt mechanism. MPC stands for multiple point clouds.

Two algorithms are considered for fusion of the acquired depth images: a point-to-plane ICP algorithm and a state-of-the-art registration algorithm TEASER++ [1] . We believe that the proposed method is a useful tool, which can be applied for creating 3D environment models using a consumer-grade 3D camera. The target application of the proposed method is in mobile robot navigation, where the local 3D environment models provided by the presented algorithm can be used for place recognition. However, the place recognition problem itself is not in the scope of this work.

2. Related Research

The aim of active vision is generally to elaborate control strategies to improve a perception task. In [2] the authors gave a broad survey of developments in active vision in robotic applications. Most of the previous work in the field of 3D scene reconstruction are designed for hand-held cameras [3–6] and rely on human instinct reactions. Such ap-

proaches are often prone to error if the scanned environment does not contain a sufficient number of distinctive features. To cope with this problem, some researches opted for automatic scene scanning. Such method was proposed by Tsai and Huang [7], where an RGB-D camera is mounted on a PTM and scans parts of a scene. Similarly, Byun and Han [8], use the same set-up to scan the entire room from its centre. Both approaches assume incremental changes in the camera pose between consecutive views in order to perform image registration using one of existing ICP methods. While that can secure a large number of images, it cannot guarantee sufficient information for successful image registration, and at the same time usually requires significant computational resources. Furthermore, to prevent the motion blur, the camera must move slowly, requiring relatively long scanning times.

Other researchers found a solution to this problem by developing strategies for selecting the NBV. The NBV planning is a central task for automated 3D reconstruction in robotics. Cameras mounted on autonomous vehicles and robot arms determine the new robot position based on the NBV. The NBV purpose is to secure capturing sufficient information for 3D scene reconstruction in as few steps as possible. An efficient NBV algorithm for 3D reconstruction of indoor scenes using active range sensing was proposed in [9]. The scanner used as a range sensor has 360° horizontal FoV, but limited vertical FoV. The NBV is given as the new position of the scanner in the observed environment. The gist of this approach was used for path planning for a mobile robotic platform designed for visual 3D reconstruction [10–12], and for 3D reconstruction in a table-top scenario [13,14]. An extensive overview of 3D reconstruction methods using 3D cameras can be found in [15].

NBV algorithms can rely on different types of data, such as volumetric (e.g., voxel grids) or surface (e.g., triangulated meshes) representations. The problem of NBV

selection for the volumetric reconstruction of an object by a mobile robot equipped with a camera is described in [16]. They choose the NBV as the view that provides the most gain in surface coverage and uncertainty reduction based on voxel occupancy likelihood. Another example of voxel representation in the NBV strategy for a geometrical primitive reconstruction is presented [17]. They opted for selection of viewpoints based on the volume of unknown region that can be expected. Exploration is terminated after observing approximately 95% of a scene. An example of surface-based NBV is proposed in [18], where the surface geometry components are determined sequentially from the eigenvectors, eigenvalues, view orientation, and the mean of the nearby points. Each surface point is processed and classified as either a core, frontier or outlier point. View proposals are generated to maximise sensor coverage of the estimated planar surface around each frontier point. The NBV is selected based on minimising total travel distance. However, methods [16] and [18] do not consider indoor scenes of simple planar geometry which lack features for reliable registration. Our algorithm addresses this problem by requiring that the NBV overlaps with one of the previous views in such a way that the overlapping scene region allows unambiguous view registration.

3. Active vision

In this section, an active vision approach for selecting the NBV is proposed. A 3D camera mounted on a PTM captures an image from the current camera position. The captured image is processed in order to select the next optimal viewing angle from the same viewpoint, or more precisely, the next optimal orientation of the PTM defined with the new pan angle α and tilt angle β. Hence, a viewing angle is described as a pair $\varphi=(\alpha, \beta)$. The set of all viewing angles from which the images were captured during the scene reconstruction process is denoted as Φ.

Every captured image must overlap with some of the other images so that the information content of the overlapping image region has sufficient information for reliable image registration. Furthermore, the next camera FoV must cover as much of the previously unexplored scene regions as possible. Consequently, the selection of the optimal next view direction is based on the following two cues:

(1) the percentage of unexplored scene regions contained in the next FoV, and
(2) the information content of the region in the image acquired from the next view which overlaps with one of the previously acquired images.

3.1. Exploring the Scene

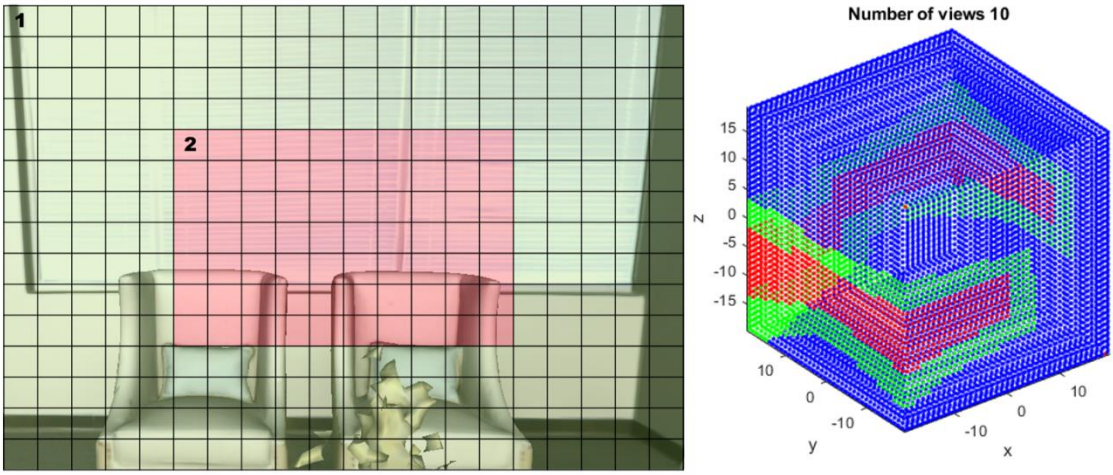
In order to determine the percentage of unexplored scene regions contained in the next FoV, all previously explored scene regions are recorded using a unit cube centred in the current viewpoint. Using a unit cube as a representation of the 360° FoV of the considered camera system allows easier computation with insignificant error compared to using a unit sphere. The unit cube's sides are divided into $\lambda_C \times \lambda_C$ square grid cells. Every cube cell corresponds to one optical ray from the camera's viewpoint. Initially, every cube cell is assigned *FoV depth index* zero which indicates that the surrounding area of the camera is completely unexplored. After an image is captured, all unit cube cell centres are projected onto the image plane. If a cube cell centre is projected inside the image, its FoV depth index is changed to a non-zero value which indicates an explored scene region.

After projecting the current FoV onto the unit cube, the algorithm uses the updated unit cube to create a list Θ of candidates θ_i for the next view. To ensure that the new FoV overlaps with a previous one, view candidates are selected so that the centre of each candidate FoV corresponds to the centre of a cube cell with FoV depth index

greater than zero. Selecting a FoV whose centre is near the centre of one of the previous FoVs is redundant for the considered active vision task. To avoid this, we opted for a strategy with two values of the FoV depth index in order to distinguish regions of interest for the NBV selection. A captured image can be divided into two concentric regions of interest: outer and inner, with assigned FoV depth indices one and two respectively, as shown in Figure 2(a). The inner region of interest represents a quarter of the total image surface, and it would give redundant information if a point within this region would be the centre of the next view image. Selecting the next view with the centre in the outer region of interest would potentially secure enough image overlap for successful image registration and cover a significant part of the unexplored scene region. Consequently, regions of greater significance for the NBV selection are assigned FoV depth index one, while regions with lesser significance are assigned FoV depth index two. The next view candidates are generated only from the cube cells with FoV depth index one.

After the next view is selected, the cube is updated. Each cube cell is projected onto the camera image. If the cell centre is contained within the image, then the cell is assigned the greater of the two values: the current cell FoV depth index and the FoV depth index corresponding to the region of interest in the image. This ultimately enables keeping a record of previously explored scene regions. An example of a unit cube with projections of ten FoVs is shown in Figure 2(b).

The algorithm estimates the amount of unexplored region which would be contained in every candidate FoV by computing the percentage of cells contained in the candidate FoV with FoV depth index equal to zero. This percentage is denoted in this paper by $\chi(\theta_i)$.



(a) (b)

Figure 2. Regions of interest for the NBV selection (a) and visualization of the projection cube and FoV depth indices assigned to the cube cells (b). The cube contains the record of ten views. Regions of a greater significance for active vision (FoV depth index = 1) are coloured green, while regions that would result in redundant views are coloured red (FoV depth index = 2). Unexplored regions (FoV depth index = 0) are coloured blue.

3.2. Overlap Information Content

By projecting the points of the *k*-th view to the image acquired from the *l*-th view, the algorithm checks which point from the *k*-th view is visible from the *l*-th view. To ensure that the overlapping region of two captured images has enough information for reliable image registration, we rely on the orientation of surfaces contained in this overlapping image region. For that purpose, each point in the image is assigned a normal representing the unit vector orthogonal to the local surface in the close neighbourhood of that point. The normal is directed from the surface outwards and indicates the orientation of the surface. In this paper, 3D points with assigned normal are referred to as oriented 3D points. Matching a pair of oriented 3D points between two views, whose normals are not parallel, provides sufficient information to determine the orientation of one view with respect to the other.

In this paper, ${}^{l}p_{m,k}$ denotes the 3D point corresponding to the pixel m of the k-th image represented in the RF of the l-th image and ${}^{l}n_{m,k}$ represents the normal of point ${}^{l}p_{m,k}$. Distribution of surface orientations in the overlapping image region of the two considered views can be represented by the covariance matrix

$$\Sigma_{kl} = \sum_{m \in O_{kl}} {}^{k}n_{m,k} \, {}^{k}n_{m,k}^{T} \tag{1}$$

where O_{kl} is the set of overlapping points between views k and l.

The eigenvector corresponding to the largest eigenvalue of this matrix represents the dominant normal direction in the overlapping image region. The eigenvector corresponding to the second largest eigenvalue represents the second dominant normal direction. If the second eigenvalue is sufficiently large, this indicates that there is a sufficient number of oriented points in at least two orthogonal directions, which would provide sufficient information for estimating all three degrees of freedom of the relative orientation between the two considered views. Hence, we define *overlap information content* (OIC) as the second-largest eigenvalue of covariance matrix Σ_{kl}.

The OIC is computed between each candidate θ_i, acquired from view candidate list Θ, and all previous views $\varphi \in \Phi$. The OIC computed for θ_i and one of the previous views φ is denoted by $\psi(\theta_i, \varphi)$. The highest of all OIC computed for a particular candidate θ_i is used to compute the candidate score.

For faster calculation of the optimal next view direction cues, the captured image is subdivided into smaller cells of $\lambda_{cell} \times \lambda_{\text{cell}}$ pixels, illustrated in Figure 2(a). Every cell is assigned a covariance matrix, calculated by (1) for all points in that cell, and the OIC is calculated by summation of the covariance matrices of the overlapping cells.

We assume that the information about the camera viewing angle is known but imprecise, which means that the true viewing angle is different from the *setpoint* camera

angle commanded to the PTM. We model the imprecision of the PTM by assuming that the true viewing angle can be obtained by rotating the camera oriented in a particular setpoint camera orientation for an angle $\leq \delta_{\max}$ around an arbitrary axis. To take into account, the imprecision of the PTM, only the image regions of the views θ_i and φ, with a high probability of being overlapped should be considered in computation of OIC. Hence, the points closer than $2\delta_{\max} f$ to the image boundary are not considered in computation of OIC, where f is the camera focal length. The remaining points represent the *active image region*. Only the cells inside this image region are considered in computation of OIC. These cells are referred to in this paper as *active cells*.

3.3. Next View Candidate Score

The following score is computed for each candidate θ_i

$$\rho(\theta_i) = w_\chi \, \chi(\theta_i) + \max_{\varphi \in \Phi} \big(\psi(\theta_i, \varphi)\big) \tag{2}$$

where w_χ is an empirically determined weight factor. Only the candidates with sufficient overlapping with one of the previous views are considered, i.e. the candidates θ_i which satisfy:

$$\max_{\varphi \in \Phi} \psi(\theta_i, \varphi) > w_\psi \tag{3}$$

where w_ψ is an experimentally determined threshold.

The calculated score implies how useful it would be to choose a specific candidate as the new camera view direction. The candidate with the highest score is selected as the new view direction.

4. View Registration

After exploring a sufficient percentage of the environment visible from the given viewpoint, the point clouds must be aligned with respect to a global coordinate system in order to obtain a complete point cloud that represents the entire scene.

There are numerous methods for solving the image registration problem based on several fundamental point set registration algorithms [19]. Consistent registration can be achieved in two ways: as bundle adjustment [20] and with global optimization [21]. The latter solves all pairwise registrations jointly, while the bundle adjustment produces jointly optimal 3D structures and viewing parameters. However, there are some solutions which combine the results of both registration methods. For example, authors in [22] use the initial registration by the ICP algorithm as initial specification for the bundle adjustment. However, analysis and comparison of different registration algorithms is out of the scope of this paper.

The algorithm described in this section performs registration of the acquired depth images using the point-to-plane ICP algorithm [23] applied in a recursive multi-view registration framework.

4.1. Determining Point Correspondences

A common approach to view registration is to minimize the distance between corresponding points of two aligned views. For successful image registration, it is crucial to determine correct point correspondences between overlapping views. A common approach for establishing point correspondences is to use keypoint detectors, which are matched between views using local descriptors. However, these approaches require textured surfaces or distinctive 3D shapes, which are not always available in indoor scenes.

In the research presented in this paper, we investigate an approach for establishing correspondences between points lying on flat featureless surfaces, which relies on the camera pose information provided by the PTM and the assumption that all images are acquired from the same viewpoint.

Let us consider the alignment of two views φ_k, $\varphi_l \in \Phi$ with the same viewpoint selected by the active vision approach described in the previous section. Furthermore, let m_k and m_l' be two image points acquired from the k-th and l-th view respectively representing the same scene point and let ${}^{l}p_{m',l}$ and ${}^{k}p_{m,k}$ be their corresponding 3D points. Assuming an ideal PTM, the positions of the point ${}^{l}p_{m',l}$ and the point ${}^{l}p_{m,k}$ would be identical, where ${}^{l}p_{m,k}$ is obtained by transforming the point ${}^{k}p_{m,k}$ to the RF of the l-th view with the rotation matrix computed from the setpoint camera angles. In reality, the imprecision of the PTM would result in certain distance between ${}^{l}p_{m',l}$ and ${}^{l}p_{m,k}$. The bounds of this distance can be estimated using the assumption about the maximum orientation error $\delta_{\max}$ introduced in the previous section. In our approach, points m_k can be associated with point m_l' only if the following condition is satisfied

$$\left\| {}^{l}p_{m',l} - {}^{l}p_{m,k} \right\| \leq d_a. \tag{4}$$

Threshold d_a is calculated as

$$d_a = 2\left\| {}^{l}p_{m,k} \right\| \delta_{max} \tag{5}$$

Analogously, the difference between the angles between the normals in corresponding points ${}^{l}p_{m',l}$ and ${}^{l}p_{m,k}$ is limited to an experimentally determined value ω, i.e.

$$arccos\left({}^{l}n_{m',l}^{T} \; {}^{l}n_{m,k} \right) \leq \omega. \tag{6}$$

Since the images are acquired approximately from the same viewpoint, the distance of

any scene point with respect to the camera is approximately the same for all views. Hence, m and m' can represent the same scene point only if their distance to the camera is approximately the same, i.e. if

$$\| {}^{l}p_{m',l} \| - \| {}^{l}p_{m,k} \| \leq r_z \tag{7}$$

where r_z is an experimentally determined distance threshold.

In order to reduce the computation time, the overlapping point set O_{kl} is sampled and for every sampled point of the k-th view a corresponding point is searched in the l-th view. Each point $p \in O_{kl}$ is assigned the *information content factor* (ICF) and the sampling is performed in such a way that the probability of selecting a point from O_{kl} is proportional to this factor. The ICF is defined by [24]:

$$w_{m,k} = \frac{1}{{}^{k}n_{m,k}^{T} \, \Sigma_{kl} \, {}^{k}n_{m,k}}. \tag{8}$$

The denominator of the ICF represents the projection of matrix Σ_{kl} in direction of the local surface normal of the point $m \in V_k$. Hence, the ICF assigned to a particular point m is the greater the smaller the number of image points in O_{kl} with a direction parallel to the normal of m. The sampling weighted with the ICF favours selection of points whose normal has a component perpendicular to the dominant normal direction in the overlapping image region between the two considered views, thereby providing sufficient information for determining the relative orientation of the two views. The result of the described sampling process is a set of points $C_{kl} \subset O_{kl}$, which represents candidates for establishing point associations between the k-th and l-th view.

Finally, every point $m \in C_{kl}$ is transformed to the l-th view and associated with the nearest point $a_m \in V_l$ inside a square window in the l-th image. The nearest point a_m is defined as

$$a_m = \underset{m' \in W_m}{argmin} \left\| {}^{l}p_{m',l} - {}^{l}p_{m,k} \right\| \quad (9)$$

where W_m is a set of points $m' \in V_l$ which are contained inside a square neighbourhood ($w_w \times w_w$) in the l-th image cantered in $f\left(P\ {}^{l}p_{m,k}\right)$ and which satisfy (4), (6) and (7). Square window size is determined as following

$$w_w = 2f\delta_{max} \quad (10)$$

where f is the camera focal length.

The result is a set A_{kl} of pairs of corresponding points (m, a_m). The next step is performing an image registration task using these correspondences.

4.2. Multi-Image Registration

Zhu et.al. [25] solved the multi-view registration problem using a spanning tree based on estimated overlap percentage between views, where each tree node represents a view. Two nodes are connected if the two corresponding views overlap sufficiently. They perform view registration recursively by registration of view pairs representing connected nodes in the spanning tree. This strategy is also used in our approach.

The first view φ_1 is taken as the reference view and all other views are aligned w.r.t its RF S_1, which represents the global RF. Neighbourhood relations are defined between the views using OIC. Two views φ_k and φ_l are neighbours if

$$\psi(\varphi_k, \varphi_l) > w_\psi \quad (11)$$

The set of all collected views Φ is subdivided into subsets Φ_i, i=1,…, n_Φ, where subset Φ_1 contains only one view φ_1. View φ_k belongs to a subset Φ_i, i=2,…, n_Φ, if $\varphi_k \notin \Phi_{i-1}$ and it has a neighbour in Φ_{i-1}. The views are aligned recursively, where each

view $\varphi_l \in \Phi_i$ is aligned with one of its neighbours $\varphi_k \in \Phi_{i-1}$. First, all neighbours of φ_1, i.e. all views from subset Φ_2, are aligned with φ_1. As the result, their pose with respect to the global RF is obtained. Next, all views which are not in Φ_1 or Φ_2, but have neighbours in Φ_2 are aligned with their neighbours from Φ_2 and their poses with respect to the global RF are obtained. This procedure continues until all views are assigned their poses with respect to the global RF. Notice that every view, except the referent view φ_1, is registered with a single neighbouring view.

The alignment of two neighbouring views is performed using a point-to-plane ICP algorithm which minimizes the following cost function

$$E(R_l, t_l) = \sum_{(m,m')\in A_{kl}} w(m,m') \left({}^{k}n_{m,k}^{T} R_k^T \left(R_l \; {}^{l}p_{m',l} + t_l - R_k^0 \; {}^{k}p_{m,k} - t_k^0 \right) \right)^2 \quad (12)$$

where rotation matrix R_l and translation vector t_l define the pose of the RF of the l-th point cloud with respect to S_1, which is optimized. Rotation matrix R_k^0 and translation vector t_k^0 represent the initial pose of the k-th view, which is not changed until the minimization is completed, and $w(m, m')$ is variable point weight depending on the uncertainty of measured point position. In the case where the information about the measurement uncertainty is available, weights $w(m, m')$ allow more accurate measurements to contribute more to the total cost function. In the case of registration of depth images acquired by a RGB-D camera, these weights can be computed using the method proposed in Section 4.3. This procedure is performed iteratively. After each iteration, the algorithm finds new correspondences A_{kl} between views.

After the last iteration, each view is aligned with one of its neighbours. To align all views with respect to the reference view, the final view poses are computed recursively by

$$T_k' = T_l' \cdot (T_l^0)^{-1} T_k \tag{13}$$

where $k > l$, T_k is the homogeneous transformation matrix defining the pose of the k-th view after the last iteration of the optimization procedure, T_k' and T_l' define the final pose of the k-th and l-th view respectively and T_l^0 represents the initial pose of the l-th view. The relation between the homogeneous transformation matrices appearing in (13) and the corresponding rotation matrix and translation vector is described by

$$T_k = \begin{bmatrix} R_k & t_k \\ 0 & 1 \end{bmatrix} \tag{14}$$

4.3. Depth Sensor Gaussian Noise

Images acquired by 3D cameras can be observed as point clouds in the Euclidean space with a defined 3D Cartesian coordinate system, referred to in this paper as the XYZ space. Registration of two views can be performed by minimizing the sum of distances between the corresponding points in the XYZ space. However, measurements obtained by real 3D cameras are affected with measurement noise, which can significantly distort images and subsequently affect the image registration task. The effect of the measurement noise can be minimised with image processing which takes into account the properties of the applied sensor, where more accurate measurements are considered with higher weights than measurements with higher uncertainty. Liu et.al. [26] introduced depth-based point weights to the general ICP algorithm to take into account the measurement noise. The point pairs with smaller Euclidean distances are used for registration, and each point pair is assigned a weight according to the depth values. The same principle is used in our method. Nevertheless, we compute the point weights differently, using the relation between the Euclidean space and the disparity space, where the measurement noise is homogeneous.

Using off-the-shelf RGB-D cameras, such as Microsoft Kinect, is cost-effective and affordable option in contrast to expensive laser scanners. Therefore, in the following sections, we consider the RGB-D camera for performing the scene reconstruction task. The measurement model of these cameras is analogous to the model of stereo vision systems addressed in [27]. In [27], it is noted that measurements obtained by such sensors represented in the XYZ space have non-homogeneous and non-isotropic noise. The measurement uncertainty of these cameras grows with the square of the distance from the camera [27]. This problem is addressed by representing the camera measurements in the disparity space, where the noise has isotropic and homogeneous behaviour. Hence, an ICP procedure which minimizes the sum of point distances in the disparity space is expected to give better results than an ICP procedure which minimizes the distances in the XYZ space.

The RGB-D camera output is a combination of an RGB image and its corresponding depth image. Consequently, an RGB-D image pixel q can be represented in the disparity space with its 2D image coordinates (u, v) and corresponding disparity (d), that is $q=[u\ v\ d]^{\mathrm{T}}$. We adopt the RGB-D camera model from [28]. Given a point $p=[x\ y\ z]^{T}$ in the XYZ space, the point in the disparity space can be computed by

$$q(p) = \begin{bmatrix} f\frac{x}{z} \\ f\frac{y}{z} \\ \kappa\left(1 - \frac{z_n}{z}\right) \end{bmatrix} \tag{15}$$

where (x, y, z) are point coordinates in the XYZ space, f is the camera focal length, z_n is the nearest distance at which the RGB-D can detect objects, and κ is the parameter of the disparity-depth relation. This representation is referred to in this paper as the UVD-space.

Correlation of geometric properties between the XYZ and the UVD space is demonstrated in [29], where it is also proven that a curve of the *n*-th order in the XYZ space corresponds to a curve with the same order in the UVD space. Hence, any planar surface in the XYZ space corresponds to a planar surface in the UVD space. Generally, any planar surface in the XYZ space can be uniquely described with its normal vector and its offset from the origin of the coordinate system. Normal $n=[n_x\ n_y\ n_z]^T$ of a planar surface in the XYZ space and normal η of the same planar surface represented in the UVD space are related by the following equation (see Appendix A)

$$\eta = \frac{1}{\sqrt{n_x^2 + n_y^2 + \left(\frac{f}{\kappa z_n} n^T p\right)^2}} \begin{bmatrix} n_x \\ n_y \\ \frac{f}{\kappa z_n} n^T p \end{bmatrix}. \tag{16}$$

Having defined point and normal in the UVD-space by (15) and (16), we can formulate the registration cost in that space by

$$E(R_l, t_l) = \sum_{(m,m')\in A_{kl}} \left({}^{k}\eta_{m,k}^T \left(q\left(R_k^T \left(R_l\ {}^{l}p_{m',l} + t_l - t_k \right) \right) - q\left({}^{k}p_{m,k} \right) \right) \right)^2 \tag{17}$$

where p_m denotes the vector of 3D coordinates of the image point *m* in the XYZ space, $q(p)$ denotes the representation of a 3D point *p* in the UVD space and η_m is the normal of the image point *m* in the UVD space. In Appendix A, it is shown that minimizing (17) is equivalent to minimizing (12) with point weights computed by

$$w(m, m') = \frac{1}{z'^2 \left(n_x^2 + n_x^2 + \left(\frac{f}{\kappa z_n} n^T p \right)^2 \right)} \tag{18}$$

where *z'* is the z-coordinate of the point $p_{m'}$.

5. Experiments

We implemented the proposed approach in C++ language and verified the discussed method by simulation. The simulation is aimed to test the proposed methods for active vision and scene reconstruction in a controlled environment.

The experiment was carried out on a publicly available dataset Replica [30] which was acquired with a custom-built RGB-D capture rig with an IR projector. We used the mesh models of nine rooms: *Hotel 0*, *Office 0-4,* and *Room 0-2*. A virtual camera is placed in the origin of the room's coordinate system which was chosen based on probable camera placement in the room. The camera movement is simulated by changing the camera view direction from the same viewpoint. The algorithm creates a depth image from a particular view using the Visualization Toolkit (VTK) [31]. In order to mimic a real RGB-D camera, as described in the previous section, we superposed the Gaussian measurement noise with zero mean and standard deviation $\sigma_d = 0.5$ to the disparity of a simulated depth sensor. According to [28], this standard deviation is a fair estimate of disparity error of the Kinect sensor.

While generating an organized point cloud from a depth image, local surface properties are computed through a normal estimation process. The normal estimation is performed using the Point Cloud Library (PCL) [32] based on the neighborhood of points within a sphere of a given radius. Since measurement noise has non-homogeneous and non-isotropic behavior in the XYZ space, this sphere radius should be variable for different point distances from the camera. To make this process simpler, we opted to perform normal estimation process in the UVD space, where the measurement noise is approximately homogeneous, so that we can use the same sphere radius for all points.

After each image acquisition, the active vision approach proposed in the third section is applied to select the NBV. This selected view is referred to in this section as

the *setpoint*. Imprecise PTM is simulated by an additional rotation of the virtual camera for a perturbation angle of 3° about a randomly selected axis before capturing a depth image. This simulates possible camera drift in the real testing conditions. This perturbed view direction is referred to in this section as the *actual view direction*.

When active vision captures n_{img} images, the algorithm proceeds to image registration task. The number of images n_{img} is specified by the user. The results of the registration procedure are referred to as the *estimated view directions*. The values of the active vision and multi-view registration algorithm parameters used in the experiments are shown in Table 1.

Table 1. The algorithm parameters used in the experiment.

ACTIVE VISION					***VIEW REGISTRATION***							
w_{img}	h_{img}	λ_C	λ_{cell}	w_χ	r_z (m)	ω (°)	δ_{max} (°)	f_c	κ[1]	w_w	w_{ROI}	w_ψ
320	240	39	16	1.0	0.03	10	3	293.172	$2.85 \cdot 10^{-6}$	62	62	0.01

The main contribution of this paper is the active vision approach, which selects the NBV according to a measure of the information useful for depth image registration contained in the overlapping image region of the next view and the previously acquired views. This is especially important in the environments with very few features, such as rooms with big blank walls or empty and long corridors. The usefulness of the proposed active vision approach can be evaluated by the success of registration of depth images acquired by this approach.

In order to do so, we additionally simulated uniform camera movement, referred to as a sequential scan, while respecting Kinect sensor field of view, and compared suc-

[1] The parameter κ is adopted from [28].

cess of image registration of images obtained with sequential scan to the images obtained with the active vision approach. To make the results comparable, we secured approximately equal percentage of remaining unexplored scene regions, as explained in subsection 3.1. Both approaches capture 30 images from the same viewpoint in the room, starting with camera optical axis parallel to the ground. As active vision approach of selecting the next view depends on OIC, the remaining unexplored scene region of each room is different, but averages approximately 11.7%, measured by the percentage of unexplored cells of the unit cube explained in Subsection 3.1. In the sequential scanning, three tilt angles of the virtual camera are used: –34°, 0° and 34°. For each tilt angle, 10 pan angles uniformly distributed in the range from 0° to 360° are used. This scanning procedure leaves 11.5% of unexplored scene region.

Success of the two described scanning approaches is measured with the proposed scene registration method, described in Section 4. We measured the accuracy of the proposed scene registration method by computing the angle-axis representation of the:

(1) estimated relative orientation between two neighbouring views,
(2) estimated absolute orientation of each view with respect to the global coordinate system.

The angle between the two compared view directions is taken as a measure of orientation estimation error and it is denoted by δ. The set of orientation estimation errors of all views of a particular scene is denoted by Δ.

The proposed multi-image registration (Section 4.2.) uses neighbourhood relations between the views based on the OIC, while the sequential scanning captures images purely based on uniform camera movement without using any information about the scene context. Therefore, additional criteria for determining neighbouring views had

to be used. We applied segment of the algorithm described in Section 3.2. with relaxed parameters, meaning that image boundary is not taken into consideration and $w_\psi = 12$. After a sequential scan is completed, the algorithm goes through all images and projects the points of the *k*-th view onto the image acquired from the *l*-th view, where $l < k$. Each view is paired with one of the previous views with which it has the largest overlap.

The results are shown in Table 2. The both approaches achieve the estimated orientation error less than the perturbation angle of 3°. Image registration applied to the images acquired with the active vision approach achieved the orientation error that is three times less compared to the sequential scanning method.

Table 2. Comparison of the active vision approach (A) with the sequential scanning (S). Success of the two described scanning approaches is measured with the image registration accuracy of the proposed scene registration method. The accuracy is shown through the errors of estimated relative orientation between two neighbouring views for multi-view registration for nine room models from the Replica dataset [30] seen from 30 views after image registration performed in 10 iterations.

ORIENTATION ERROR									
Orientation error		**Relative**				**Absolute**			
		$\max_{\delta_i \in \Delta} \delta_i$ [°]		$\overline{\delta}$ [°]		$\max_{\delta_i \in \Delta} \delta_i$ [°]		$\overline{\delta}$ [°]	
Scanning method		**S**	**A**	**S**	**A**	**S**	**A**	**S**	**A**
ROOM	Hotel 0	**1.71**	1.90	**0.24**	0.31	**1.49**	1.94	**0.55**	1.24
	Office 0	7.29	**0.38**	0.77	**0.16**	7.88	**0.95**	2.54	**0.62**
	Office 1	7.07	**0.36**	0.59	**0.16**	9.89	**0.67**	1.57	**0.30**
	Office 2	2.66	**0.78**	0.42	**0.18**	2.23	**0.76**	0.88	**0.32**
	Office 3	4.13	**1.14**	0.48	**0.25**	4.35	**1.43**	1.60	**1.14**
	Office 4	13.86	**0.32**	1.50	**0.12**	17.28	**0.58**	3.93	**0.27**
	Room 0	4.08	**0.80**	0.50	**0.22**	3.25	**1.03**	1.43	**0.75**
	Room 1	5.99	**0.41**	0.41	**0.16**	6.54	**0.86**	0.95	**0.35**
	Room 2	7.61	**0.42**	0.78	**0.14**	7.68	**0.76**	2.99	**0.52**
AVERAGE ERROR PER ALL ROOMS		6.04	**0.72**	0.63	**0.19**	6.73	**1.00**	1.83	**0.61**

Furthermore, it can be noticed that the active vision method achieves significantly better results on eight of nine rooms, while on room *Hotel 0*, the sequential scanning has slightly better result than the active vision. This is a smaller room with many

distinctive 3D surfaces. In the other rooms, dominated by textureless surfaces, the advantage of the proposed active vision strategy is obvious.

Additionally, we evaluated our method by measuring the accuracy of the camera orientation estimation obtained by two registration methods: point-to-plane ICP and TEASER++ [1]. In order to evaluate the contribution of the OIC cue in the criterion for selection of the NBV, the performance of the proposed approach is compared to a modified version of our algorithm, which does not use OIC in the criterion (2). Instead of eliminating view candidates according to criterion (3), based on OIC, a simpler elimination criterion is used. This criterion requires that at least 25% of the active cells of the next view candidate overlap with one of previously acquired views. Active cells are defined in Section 3.2.

The evaluation was conducted by capturing 30 images and performing the multi-view registration proposed in Section 4 with ten iterations. To demonstrate the contribution of each particular component of the proposed method, we ran experiments in four different set-ups:

(1) without contribution of the OIC cue, computing normals directly in the XYZ space; w=1 in (12),
(2) with contribution of the OIC cue, computing normals directly in the XYZ space; w=1 in (12),
(3) computing normals in the UVD space; w=1 in (12),
(4) computing normals in the UVD space; w in (12) computed by (18).

Quantitative results of the proposed approaches are shown in Table 3 and Table 4 by average relative orientation estimation error $\overline{\delta}$, and average absolute orientation estimation error respectively, after 10 iterations and maximal error in the set of 30 views for all rooms.

Table 3. The errors of estimated relative orientation between two neighbouring views, for the proposed approach for active vision and multi-view registration for nine room models from the Replica dataset [30] seen from 30 views after image registration performed in 10 iterations.

RELATIVE ORIENTATION ERROR										
OIC cue			OFF		ON		ON		ON	
Geometric space			XYZ		XYZ		UVD		UVD	
Point weights computed by (18)			OFF		OFF		OFF		ON	
Registration method			**T++**	**ICP**	**T++**	**ICP**	**T++**	**ICP**	**T++**	**ICP**
ROOM	Hotel 0	$\max_{\delta_i \in \Delta} \delta_i$ [°]	2.24	3.13	3.18	1.74	1.84	6.01	1.84	1.90
	Office 0		4.87	9.61	1.09	1.02	2.47	1.59	2.47	0.38
	Office 1		4.99	1.53	0.56	0.27	0.91	0.66	0.91	0.36
	Office 2		17.04	12.91	4.15	2.57	3.86	2.91	3.86	0.78
	Office 3		7.83	9.80	2.85	2.56	1.39	41.15	1.39	1.14
	Office 4		34.25	36.73	2.29	3.14	1.01	0.68	1.01	0.32
	Room 0		1.90	3.50	3.66	3.41	0.83	0.69	0.83	0.80
	Room 1		2.31	8.16	1.98	1.07	0.92	2.05	0.92	0.41
	Room 2		125.62	121.35	1.98	3.57	1.49	1.28	1.49	0.42
AVERAGE ERROR PER ALL ROOMS			**22.34**	22.97	2.42	**2.15**	**1.64**	6.34	1.64	**0.72**
ROOM	Hotel 0	$\overline{\delta}$ [°]	0.43	0.45	0.41	0.24	0.60	0.43	0.60	0.31
	Office 0		0.51	0.56	0.31	0.23	0.49	0.22	0.49	0.16
	Office 1		0.46	0.20	0.27	0.12	0.38	0.21	0.38	0.16
	Office 2		1.46	1.37	0.67	0.41	0.56	0.31	0.56	0.18
	Office 3		0.75	0.62	0.49	0.34	0.42	1.69	0.42	0.25
	Office 4		2.53	2.46	0.50	0.41	0.38	0.23	0.38	0.12
	Room 0		0.54	0.42	0.53	0.48	0.44	0.26	0.44	0.22
	Room 1		0.43	0.51	0.36	0.17	0.36	0.26	0.36	0.16
	Room 2		6.01	5.81	0.46	0.38	0.30	0.21	0.30	0.14
AVERAGE ERROR PER ALL ROOMS			1.46	**1.38**	0.44	**0.31**	0.44	**0.43**	0.44	**0.19**

The presented analysis demonstrates that including OIC in the NBV selection criterion significantly contributes to the registration accuracy. Furthermore, the results indicate that performing normal estimation in the UVD space results in lower orienta-

tion estimation error. Further significant improvement is evident in the fourth experiment set-up, where the point weights were introduced. The average of maximal absolute orientation estimation errors of all rooms in that case is 1.00°. Moreover, maximal δ does not exceed the initial error of 3°, while average $\overline{\delta}$ for all rooms is less than 1°.

Table 4. The errors of estimated absolute orientation of each view with respect to the global coordinate system, for the proposed approach for active vision and multi-view registration for nine room models from the Replica dataset [30] seen from 30 views after image registration performed in 10 iterations.

ABSOLUTE ORIENTATION ERROR										
OIC cue			OFF		ON		ON		ON	
Geometric space			XYZ		XYZ		UVD		UVD	
Point weights computed by (18)			OFF		OFF		OFF		ON	
Registration method			**T++**	**ICP**	**T++**	**ICP**	**T++**	**ICP**	**T++**	**ICP**
ROOM	Hotel 0	$\max_{\delta_i \in \Delta} \delta_i$ [°]	2.24	3.23	4.09	2.70	3.72	6.10	3.72	1.94
	Office 0		4.89	9.56	1.59	1.13	3.27	2.00	3.27	0.95
	Office 1		5.74	1.26	1.60	0.35	1.17	1.20	1.17	0.67
	Office 2		17.90	14.93	4.92	4.73	3.57	3.42	3.57	0.76
	Office 3		10.03	10.03	2.80	2.35	1.90	42.56	1.90	1.43
	Office 4		56.71	51.04	2.67	3.72	1.94	1.40	1.94	0.58
	Room 0		1.95	4.36	3.73	3.41	3.57	1.72	3.57	1.03
	Room 1		4.31	8.42	3.93	1.54	2.09	2.08	2.09	0.86
	Room 2		125.37	125.75	4.69	5.34	1.49	1.73	1.49	0.76
AVERAGE ERROR PER ALL ROOMS			25.46	**25.40**	3.34	**2.81**	**2.52**	6.91	2.52	**1.00**
ROOM	Hotel 0	$\overline{\delta}$ [°]	1.08	1.90	3.25	1.97	1.65	2.18	1.65	1.24
	Office 0		1.31	1.95	0.92	0.70	1.91	1.61	1.91	0.62
	Office 1		0.98	0.38	0.83	0.19	0.61	0.73	0.61	0.30
	Office 2		3.69	2.95	2.62	1.49	1.15	1.81	1.15	0.32
	Office 3		2.02	1.76	1.00	0.77	1.18	19.23	1.18	1.14
	Office 4		5.61	5.45	2.01	1.27	1.24	0.69	1.24	0.27
	Room 0		1.05	1.65	2.34	2.76	1.62	0.96	1.62	0.75
	Room 1		1.21	2.44	1.08	0.49	1.17	0.74	1.17	0.35
	Room 2		52.44	53.56	1.68	2.14	0.82	1.22	0.82	0.52
AVERAGE ERROR PER ALL ROOMS			**7.71**	8.00	1.75	**1.31**	**1.26**	3.24	1.26	**0.61**

According to the presented analysis, the point-to-plane ICP has better accuracy than TEASER++ in the given conditions. This can be explained by the fact that the point-to-plane distance is insensitive to the point position on a plane, i.e. it requires only that two corresponding points lie on the same plane.

On the other hand, the point-to-point distance minimized by the TEASER ++ requires more precise correspondences. The advantage of the TEASER ++ over ICP algorithms is that it does not require initial relative pose of the registered scans. However, in the case considered in this paper this advantage does not have effect because initial relative orientations are provided by the PTM.

In Figure 3, qualitative results of the proposed scene reconstruction method are presented, where the obtained 3D room model before and after image registration is visualized. It can be noticed that the applied image registration procedure successfully corrected original misalignments of the point clouds.

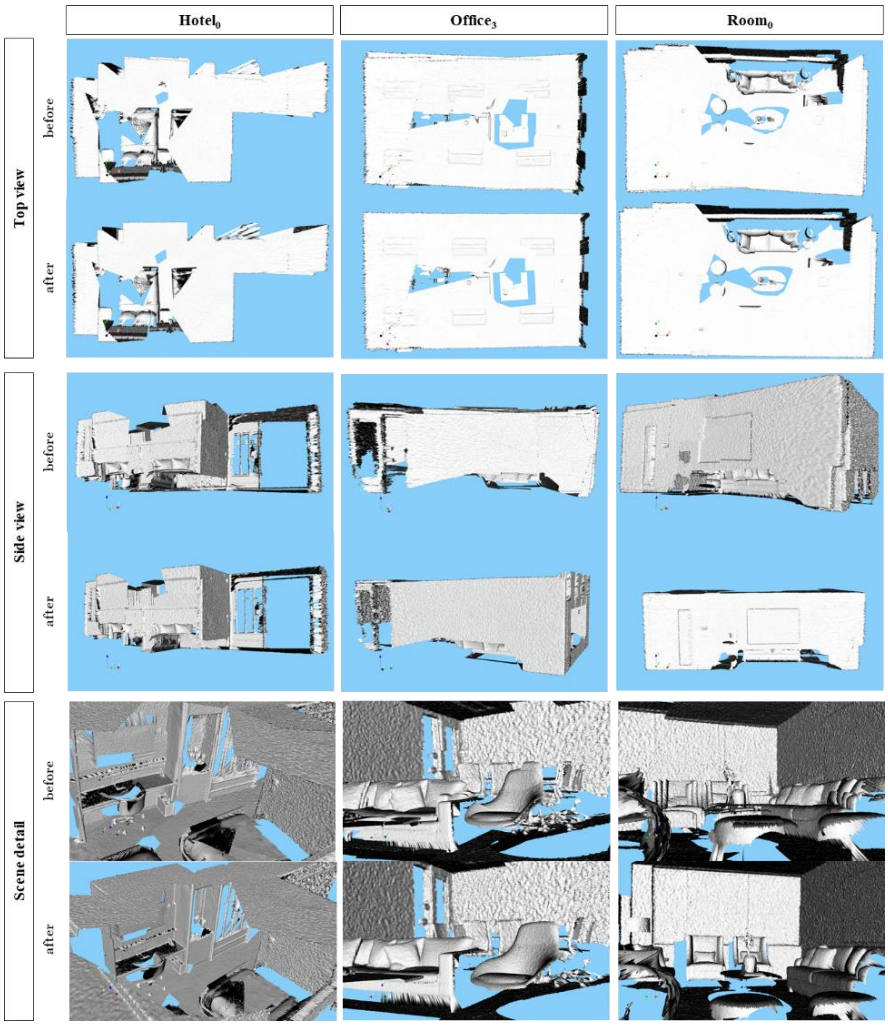


Figure 3. Three reconstructed rooms: *Hotel 0*, *Office 3* and *Room 0*, before and after scene registration performed in 10 iterations.

Figure 4 visualizes the camera movement using the unit cube to illustrate the distribution of the captured images in different rooms. The colours correspond to FoV depth indices defined in Subsection 3.1. The presented results show how the automated camera movement is directly related to each room individually. This confirms that view selection is correlated to the room features and camera placement in each room.

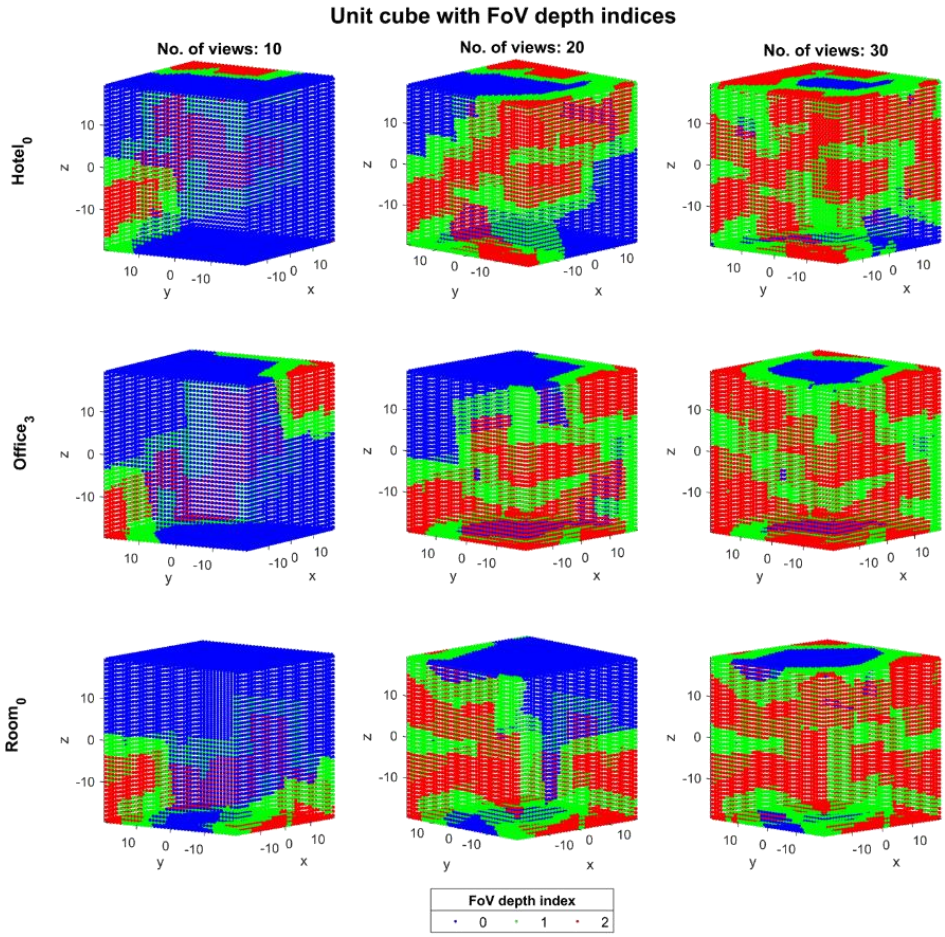


Figure 4. Visualization of the recorded views in the active vision task for *Hotel 0*, *Office 3*, and *Room 0*.

As explained in Section 3.1., the active vision algorithm selects the NBV based on previously explored scene. Consequently, with every captured scene there are more candidates for the active vision algorithm to process which requires more computation time. The average computation time of the NBV selection for nine rooms in relation to number of captured images is shown in Figure 5.

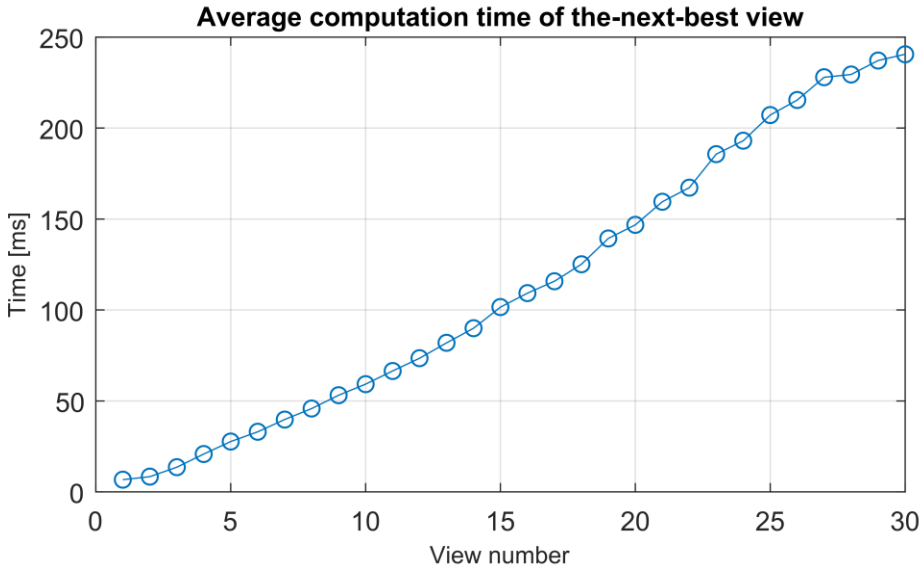


Figure 5. Visualization of the average processing time of the NBV computation in relation to number of captured images.

6. Conclusion

In this paper, we propose a NBV selection criterion which provides depth images that can be accurately registered. This active vision approach allows realization of fully automatized creation of 3D models of indoor environments using a 3D sensor mounted on a PTM. We tested the proposed approach by simulating active vision with virtual camera positioned within nine rooms from the Replica dataset [30], with synthetic measurement noise. The proposed method is evaluated by measuring the accuracy of alignment of depth images acquired by the proposed active vision strategy, using a point-to-plane ICP and the TEASER++ registration algorithm. It is demonstrated that the proposed overlap information content criterion for selection of the NBV significantly contributes to the registration accuracy.

The proposed method resulted in average orientation estimation error below 1° on a set of nine rooms. Although the proposed method is designed to correct relatively

small orientation errors caused by imprecision of PTM, this correction is necessary if we want to obtain accurate scene models.

Our future research will focus on application of the described method using real off-the-shelf RGB-D cameras mounted on a PTM.

Acknowledgements. This research has been partially supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS) and partially by the Croatian Science Foundation under the project number IP-2014-09-3155.

References

1. Yang H, Shi J, Carlone L. TEASER: Fast and Certifiable Point Cloud Registration. 2020;(c):1–42.
2. Chen S, Li Y, Kwok NM. Active vision in robotic systems: A survey of recent developments. Int J Rob Res. 2011 Sep 22;30(11):1343–77.
3. Izadi S, Kim D, Hilliges O, Molyneaux D, Newcombe R, Kohli P, et al. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In: UIST'11 - Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. 2011.
4. Whelan T, Salas-Moreno RF, Glocker B, Davison AJ, Leutenegger S. ElasticFusion: Real-time dense SLAM and light source estimation. In: International Journal of Robotics Research. 2016.
5. Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, et al. KinectFusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2011. 2011. p. 127–36.
6. Dai A, Nießner M, Zollhöfer M, Izadi S, Theobalt C. BundleFusion. ACM Trans Graph. 2017 Jul 20;36(4):1.
7. Tsai CY, Huang CH. Indoor scene point cloud registration algorithm based on RGB-D camera calibration. Sensors (Switzerland). 2017;17(8).

8. Byun J-H, Han T-D. Fast and Accurate Reconstruction of Pan-Tilt RGB-D Scans via Axis Bound Registration. 2018 Dec 1;

9. Low K-L, Lastra A. An Adaptive Hierarchical Next-Best-View Algorithm for 3D Reconstruction of Indoor Scenes. Proc 14th Pacific Conf Comput Graph Appl. 2006;

10. Dunn E, Van Den Berg J, Frahm JM. Developing visual sensing strategies through next best view planning. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009. 2009. p. 4001–8.

11. Elzaiady ME, Elnagar A. Next-best-view planning for environment exploration and 3D model construction. In: 2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017. Institute of Electrical and Electronics Engineers Inc.; 2018. p. 745–50.

12. Prieto SA, Quintana B, Adán A, Vázquez AS. As-is building-structure reconstruction from a probabilistic next best scan approach. Rob Auton Syst. 2017 Aug 1;94:186–207.

13. Monica R, Aleotti J. Surfel-based next best view planning. IEEE Robot Autom Lett. 2018;

14. Monica R, Aleotti J. Contour-based next-best view planning from point cloud segmentation of unknown objects. Auton Robots. 2018;

15. Zollhöfer M, Stotko P, Görlitz A, Theobalt C, Nießner M, Klein R, et al. State of the art on 3D reconstruction with RGB-D cameras. Comput Graph Forum. 2018;

16. Isler S, Sabzevari R, Delmerico J, Scaramuzza D. An information gain formulation for active volumetric 3D reconstruction. In: Proceedings - IEEE International Conference on Robotics and Automation. 2016.

17. Marchand É, Chaumette F. Active vision for complete scene reconstruction and exploration. IEEE Trans Pattern Anal Mach Intell. 1999;21(1):65–72.

18. Border R, Gammell JD, Newman P. Surface Edge Explorer (see): Planning next best views directly from 3D observations. In: Proceedings - IEEE International Conference on Robotics and Automation. Institute of Electrical and Electronics Engineers Inc.; 2018. p. 6116–23.

19. Zhu H, Guo B, Zou K, Li Y, Yuen KV, Mihaylova L, et al. A review of point set registration: From pairwise registration to groupwise registration. Sensors (Switzerland). 2019.

20. Triggs B, McLauchlan PF, Hartley RI, Fitzgibbon AW. Bundle Adjustment — A Modern Synthesis Vision Algorithms: Theory and Practice. Vis Algorithms Theory Pract. 2000;

21. Calvin JM, Gotsman CJ, Zheng C. Global optimization for image registration. AIP Conf Proc. 2019;2070(March).

22. Nagara K, Fuse T. Development of Integration and Adjustment Method for Sequential Range Images. Int Arch Photogramm Remote Sens Spat Inf Sci - ISPRS Arch. 2015;40(4W5):177–81.

23. Chen Y, Medioni G. Object modeling by registration of multiple range images. In: Proceedings - IEEE International Conference on Robotics and Automation. 1991.

24. Cupec R, Nyarko EK, Filko D, Kitanov A, Petrovi I. Place recognition based on matching of planar surfaces and line segments. Int J Rob Res. 2015;

25. Zhu J, Zhu L, Jiang Z, Bai X, Li Z, Wang L. Local to global registration of multi-view range scans using spanning tree. Comput Electr Eng. 2017;58:477–88.

26. Liu S, Gao D, Wang P, Guo X, Xu J, Liu DX. A depth-based weighted point cloud registration for indoor scene. Sensors (Switzerland). 2018;18(11):1–11.

27. Demirdjian D, Darrell T. Motion estimation from disparity images. In: Proceedings of the IEEE International Conference on Computer Vision. 2001.

28. Khoshelham K, Elberink SO. Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors. 2012;

29. Zhong Y, Dai Y, Li H. 3D Geometry-Aware Semantic Labeling of Outdoor Street Scenes. In: Proceedings - International Conference on Pattern Recognition. 2018.

30. Straub J, Whelan T, Ma L, Chen Y, Wijmans E, Green S, et al. The Replica Dataset: A Digital Replica of Indoor Spaces. arXiv Prepr arXiv190605797. 2019;

31. Schroeder W, Martin K, Lorensen B, Kitware I. The visualization toolkit : an

object-oriented approach to 3D graphics. Kitware; 2006. 512 p.

32. Rusu RB, Cousins S. 3D is here: Point Cloud Library (PCL). In: Proceedings - IEEE International Conference on Robotics and Automation. 2011.

Appendix A

Let $p=[x\ y\ z]^T$ be a point represented in the XYZ space and let $n=[n_x\ n_y\ n_z]^T$ be the normal of that point. Equation

$$n^T\left(p'-p\right)=0\,, \tag{19}$$

defines a plane in the XYZ space which contains point p and whose normal is n. Let's define the *signed point-to-plane distance* between p and p' as

$$e\left(p,p'\right)=n^T\left(p'-p\right). \tag{20}$$

If $z' \neq 0$, then

$$\frac{f}{z'}e\left(p,p'\right)=n_x\left(\frac{f}{z'}x'-\frac{f}{z'}x\right)+n_y\left(\frac{f}{z'}y'-\frac{f}{z'}y\right)+n_z\left(f-\frac{f}{z'}z\right). \tag{21}$$

By simple algebraic transformations of the right side of (21), the following equation can be obtained

$$\frac{f}{z'}e\left(p,p'\right)=n_x\left(\frac{f}{z'}x'-\frac{f}{z}x\right)+n_y\left(\frac{f}{z'}y'-\frac{f}{z}y\right)+\frac{fn^Tp}{\kappa z_{\mathrm{n}}}\left(\kappa\left(1-\frac{z_{\mathrm{n}}}{z'}\right)-\kappa\left(1-\frac{z_{\mathrm{n}}}{z}\right)\right) \tag{22}$$

According to the definition of the UVD space (15), equation (22) can be written as

$$\frac{f}{z'}e\left(p,p'\right)=\sqrt{n_x^2+n_y^2+\left(fn^Tp/\left(\kappa z_{\mathrm{n}}\right)\right)^2}\,\eta^T\left(q\left(p'\right)-q\left(p\right)\right), \tag{23}$$

where η is defined by (16).

If p' is any point lying on the plane defined by (19), then $e(p, p') = 0$. In that case, from (23) it follows that

$$\eta^T\left(q(p')-q(p)\right)=0\,. \tag{24}$$

Since (24) defines a plane in the UVD space, it can be concluded that the plane defined by (24) is the UVD representation of the plane defined by (19). Consequently, normal *n* in the XYZ space is represented by vector η in the UVD space.

Furthermore, $\eta^T \left(q(p') - q(p)\right)$ represents the signed point-to-plane distance between the UVD representations of points *p* and *p'*. From (23), the following relation between the point-to-plane distances in the XYZ and UVD space can be obtained

$$\eta^T \left(q\left(p'\right) - q\left(p\right)\right) = \frac{f}{z'\sqrt{n_x^2 + n_y^2 + \left(fn^T p / \left(\kappa z_{\mathrm{n}}\right)\right)^2}} e\left(p, p'\right). \qquad (25)$$

From this relation it follows that minimization of (17) is equivalent to minimization of (12), where the weight *w* is defined by (18).