Model-based pose estimation for visual indoor progress monitoring using line features

Christoper Kropp1, Christian Koch2 Markus König3,

1) Christopher Kropp, Chair of Computing in Engineering, Faculty of Civil and Environmental Engineering, Ruhr-Universität Bochum, Germany. Email: christopher.kropp@rub.de
2) Christian Koch, Centre for Structural Engineering and Informatics, Faculty of Engineering, The University of Nottingham, United Kingdom. Email: christian.koch@nottingham.ac.uk
3) Markus König, Chair of Computing in Engineering, Faculty of Civil and Environmental Engineering, Ruhr-Universität Bochum, Germany. Email: koenig@inf.bi.rub.de

Abstract: Vision-based approaches have shown to be a suitable means for deriving information from schedule-loaded Building Information Modelling (4D BIM) models in outdoor environments knowing the pose of the camera. Just like in exterior environments, BIM-registered image sequences of interior finishing works are a promising information source for the purpose of progress monitoring as well. The registration of images provides the pose of the sensor for each image frame. This allows for the direct comparison with 4D BIM models, which includes i.e. the deployment of efficient and sophisticated vision-based recognition algorithms or the visual projection of building elements augmenting the view.

However, the changing structure of the interior of a building under construction leads to different requirements and challenges for the use of vision-based methods. Methods that succeed outdoors cannot be transferred and applied indoors without any adaptation to the special needs of the indoor environment. Considering this, new approaches need to be developed in order to make pose estimation available inside buildings.

Within this paper, an overview of present vision-based methods that reveal the camera path with the pose of each frame is given. Crucial characteristics of the conduction of those methods in interior environments for the construction phase are presented in detail. Furthermore, the abilities and constraints of the vision-based methods are examined for applicability indoors. Therefore, a catalogue of requirements is established that needs to be satisfied.

According to these findings regarding the situation during interior finishing, a concept is proposed that unites the necessary features for indoor progress monitoring in a single approach. Finally, the realization of the approach is briefly addressed and preliminary results are illustrated.

Keywords: 4D BIM, BIM-registered images, indoor progress monitoring, vision-based localization.

1. INTRODUCTION

During the construction phase, deviations to the as-planned schedule can lead to delays with timely and financial losses. Re-active scheduling is an instrument that can reduce the influence of delays on subsequent construction tasks. However, this requires knowledge of the actual state of a building. In order to assess the actual state frequently, activities need to be observed. Currently, the assessment and registration of the actual state is mostly carried out manually. Especially, monitoring interior finishing processes is a very tedious activity due to structural characteristics. Increasing the degree of automation for progress monitoring allows for the necessary high frequency of inspections and steady awareness of the current state.

In recent years, the research community discovered and extensively investigated vision-based methods for state recognition related to constructional purposes. Leveraging images registered in terms of their position and orientation to schedule-loaded Building Information Modelling (4D BIM) enables sophisticated state recognition. For outdoor environments, current approaches assess photos taken from different positions around the building (Dai, Rashidi, Brilakis, & Vela, 2013; Golparvar-Fard, Peña-Mora, & Savarese, 2012; C. Kim, Son, & Kim, 2011; H. Kim & Kano, 2008). For the registration to 4D BIM, they need to be interrelated. However, the interrelation of single images in indoor environments is already infeasible with images taken from adjacent, but structurally separated rooms.

With the aim to obtain a continuous interrelation for registration with 4D BIM, this paper introduces vision-based localization with image sequences from videos for interior progress monitoring and proposes a reasonable method of choice for this challenging task.
2. BACKGROUND

In this section, an overview of the current achievements of different vision-based methods that estimate camera motion through a scene is given in order to argument for the method of choice for achieving the aim of indoor BIM-registration of image sequences. Some may reason that the following approaches of Visual Odometry (VO) and Visual Simultaneously Localization and Mapping (VSLAM) are special cases of Structure from Motion (SFM). However, since research spreads into specialized branches that have distinct assumptions, the differentiation in this section is substantiated.

2.1 Structure from Motion

SFM is a method that reconstructs the camera motion as well as revealing the observed structure of the scene. Thereby, triangulation is a crucial requirement for 3D reconstruction. This is achieved by estimating the camera motion between frames for two or more images (Richard Hartley & Zisserman, 2004). SFM can also be applied on ordered image sets or even sequences which usually observe a centered object that should be reconstructed (see Fig. 1). On the one hand, the matching of frame-to-frame image features is more challenging and more expensive due to an possibly unknown and wide baseline between the images. Even intrinsic parameters of the camera system that acquires the image data is detectable by SFM. One key that led to an automatic SFM pipeline was the introduction of advanced feature descriptors, like SIFT (Lowe, 2004) or SURF (Bay, Tuytelaars, & Van Gool, 2006). They allow for the robust matching of unordered image sets, since they provide invariance in terms of scale and rotation.

On the other hand, the usually wide baseline between images aids the accurate reconstruction of the observed object. By a final global optimization procedure with bundle adjustment, regarding the minimization of the reprojection error, the whole model is adjusted. This optimization step is computationally expensive and grows exponentially with the amount of observations. One solution to reduce the computational effort for this optimization is to use sparse methods like (Lourakis & Argyros, 2009). Approaches with respect to the whole automatic SFM pipeline are widely used (Agarwal, Snavely, Simon, Seitz, & Szeliski, 2009; Wu, 2013). Incremental approaches, like in (Wu, 2013), aim to incrementally add images to the structure in linear computation time. Other approaches deploy hierarchical cluster trees that perform local computations of coherent regions and afterwards sew the interfaces of the local computations to create one model (Farenzena, Fusiello, & Gherardi, 2009).

2.2 Visual Odometry

VO was introduced by (Nistér, Naroditsky, & Bergen, 2004) and described as an algorithm that estimates the path of a camera that is providing an image sequence system through its environment. It considers two variants of the algorithm, one for a monocular and another for a binocular system. The monocular variant of the algorithm tracks image features over frames of an image sequence. The relative pose of three frames is estimated with corresponding observations in all three images. A triangulation between the first and the third image is performed to obtain 3D information for the 2D feature points. The frames that are used to reveal 3D information are called key frames. Afterwards, the second frame is inserted into the obtained 3D coordinate system of the first and third frame to determine the right scale. This is necessary, since the translation vector between two frames of a monocular system is only definable up to a scale. Succeeding frames are iteratively added to the initial 3D coordinate system and more triangulated points are appended. The stereo variant of the algorithm can leverage the obtained 3D information directly derivable from a known stereo system baseline. Thus, information about the structure is used as long as a corresponding observation appears in a succeeding image. After the track of a feature is lost, i.e. occlusion occurs or the field of view is exited, its 3D information is discarded (Engel, Sturm, & Cremers, 2013). Neglecting past observations and derived information allows deployment in real-time applications. Although previous work has been presented using binocular systems for egomotion estimation or self-localization, the initial paper on (monocular) VO and the applied techniques allowed for real-time use at a large scale and started
a vast amount of research on that issue (Scaramuzza & Fraundorfer, 2011). Royer, Lhuillier, Dhome, & Lavest (2007) try to decrease the uncertainty in depth estimation for image features. In their approach they select frames as key frames with a maximum amount of frames in-between the sequence still sharing common features. This approach assumes that each frame contributes to a larger baseline aiding the accuracy of the triangulation process. The approach presented in Comport, Malis, & Rives (2007) avoids using the triangulation step leveraging image-to-image relations by applying quadrifocal tensors on a binocular camera system. Hence, structure of the scene is not computed primarily. For monocular systems, this is possible by using trifocal tensors. For points, seven correspondences in three images are needed (Richard Hartley & Zisserman, 2004). For lines, thirteen correspondences are engaged (Weng, Huang, & Ahuja, 1992). Instead of extracting image features, approaches like in Engel et al. (2013) are able to directly use image intensities and are referred to as direct or appearance based methods. In order to reduce the drift that occurs during the incremental addition of new images, some approaches apply regional optimization with bundle adjustment on a window of the last couple of frames (Niko, Konolige, Lacroix, & Protzel, 2005).

2.3 Visual simultaneously localization and mapping

VSLAM is a vision-based method that represents a branch considering both estimation of the camera location whilst building a map. With this technique, a global map is built that consists of landmarks which represent the estimated position of feature correspondences in a row of image frames. The information that is stored in the map allows for active search of feature recognition. Since landmarks are kept in the map despite they are not visible (i.e. not in the field of view or simply not recognized yet), they can be recognized in future image frames. Some references also call VSLAM a real-time or online version of SFM (Strasdat, Montiel, & Davison, 2010) due to the fact that the structure of the environment is kept in the map. Applications in SFM often aim for reconstruction of objects, whereas VSLAM keeps the landmarks in the map for the purpose of recognition of the same pose, i.e. for loop-closures of the path. Hence, VSLAM only needs to involve as many features as necessary to estimate the localization, which can end up in very sparsely populated maps.
methods. Filtering-based methods focus on the newest camera pose in the map. The experience of guessed landmarks and former poses is coded in the probabilistic model. The model is created and updated with every measurement of the incoming image frame. The current state of the model allows estimating the new pose for the incoming image frame and its measurements and can update the information in the model. In (Gee & Mayol-Cuevas, 2006), an approach is presented that is able to integrate an existing computer aided design (CAD) model into the filter model. This approach aims to increase robustness to partial occlusions of lines and can potentially be utilized to discover structure on a degree that is not modelled in advance. The 2D lines are parametrized by their pixel in the middle of the line, orientation in degrees and half length of the line and matched to the 3D line by a nearest neighbor matching.

Key frame-based methods focus on the direct interrelation between two selected key frames. Image frames between the key frames are used for incremental progress, but are not considered in creating the model. Furthermore, the focus is not on the recognition of landmarks in the images, but more on tracking them over a sequence of frames, which is also referred to as parallel tracking and mapping (PTAM) (G. Klein & Murray, 2007). Optimization is performed on the last couple of key frames of the sequence or even the whole sequence. This has a strong correlation to regionally optimized VO with the difference that here, a map is created. A feature once not further tracked is updated by optimization, rather than by new measurements. Using only the key frames and the currently tracked features for updates make key frame-based approaches very fast. Key frame-based methods are quite precise with a high amount of features (Strasdat et al., 2010). A recently presented approach takes advantage of appearance-based tracking to extract a high amount of measurements of a monocular frame (Engel, Sch, & Cremers, 2014). In order to enable loop-closure on key frame methods, a bag of words approach with advanced feature descriptors can be used, that permits recognition of a scene once visited (Engel et al., 2014).

It is reported that VSLAM performs well with rotations for both methods, when some translation occurs and a certain number of features is trackable (Strasdat et al., 2010), whereas the performance fails in rotation-dominant scenes, particularly in the case of key frame-based methods (Pirchheim, Schmalstieg, & Reitmayr, 2013). Key frame based methods tend to lose scale if low parallax is induced onto a tracked feature and features leave the field of view quickly, like in rotation-dominant environments, so that key frames are forced to be positioned at a high frequency. Hence, the whole process of creating a consistent map und thus reconstructing the motion stops. Filter-based methods underlie a more robust motion model that constraints motion by a filter.

### 2.4 Model-based localization

There are some approaches that localize an agent in a scene that is known while model information is available (Aider, Hoppenot, & Colle, 2005; Kitanov, Bisevac, & Petrovic, 2007; Zhu, Zheng, & Yuan, 2011). Thereby, line correspondences between the 2D image and the 3D model (Aider et al., 2005; Kitanov et al., 2007) or other maps, for example represented by grid cells (Zhu et al., 2011), are incorporated.

### 3. ANALYSIS

Indoor environments during the time of construction of a building project are very challenging for computer vision methods. Several characteristics need to be considered and analyzed in order to determine a proper method for the aim of creating a proper estimation of the camera localization within the building. In this section, several challenges that are based on the characteristics for indoor motion estimation are investigated. An analysis of these characteristics considering different aspects of vision-based methods is conducted. For indoor environments, BIM models can be utilized to achieve registered motion information. Therefore, the information from both the building model included in BIM and the image sequences need to be joined by a vision-based method.

![Example image frames out of the same sequence demonstrating the versatile set of challenges in indoor scenes under construction](image)

**Motion within the coordinate system of the building model:** In the end, motion information is only useful if it is registered with the building model. The computed scale in vision-based methods is usually not connected to real world metrics due to the lack of 3D depth information. Thus, the motion scale from frame to frame is not assignable to the overall motion scale of the building model. Furthermore, the coordinate system origins are not
the same. Vision-based methods usually refer to a world coordinate system that has its origin in the camera origin of the first image frame. But it is obvious that the first image frame does not always start at the origin of the building model coordinate system. This information must be determinable. Methods that directly work with absolute coordinates, such as methods that use RGB-D or binocular vision, are advantageous concerning that issue at first glance, since the scale of the motion is already known. Nonetheless, the registration of the coordinate system must to be performed. Contrarily, monocular systems additionally need to be scaled into the right metrics of the motion within the 3D world of the building model. Moreover, if correspondences can be extracted out of the building model (i.e. lines or planes) and be recognized in the image, these correspondences are able to solve the whole issue of scaling. In addition, uncertainties concerning the 3D structure that monocular approaches lack are turned into a certain state. Here, filter-based VSLAM methods are predestined for that task and can be revisited as it decreases the drift. Also, line-based SFM could also be an option, since computation of the path is not necessarily needed, as online and VO can be initialized with these line as a start, but cannot recognize them once the tracking is lost. Model-based localization is totally based on the provided geometry of the building model and is able to absolutely orient according to it, if correspondences in the image can be found. In this aspect, this approach is a well-suited option. Extracting points out of the building model and recognizing them is challenging and not further considered.

**Data of building model not visible:** Although the geometry of the building is available, when the building model can be accessed, it cannot be assumed that enough geometry is visible in the field of view in every frame in order to derive the camera pose (see Fig. 4a-b). Thus, the concept that is to be developed for motion estimation needs to consider these circumstances and must be able to robustly acquire and use new information from the image data that was not previously modeled for the building. This is the aspect where visual localization-only methods lack feasibility in this environment. All other vision-based method mentioned in the background section are by definition able to gather and process more information from the image sensor to estimate the motion within the building.

**Rotation-dominating motion supported:** Traversing the camera through the interior of a building and observing objects inherently means that the camera does not follow a straight path through the structure of the building. In most robotic indoor examples for the proposed vision-based algorithms, a vehicle is trying to find its path along aisles. Within the application of progress monitoring of the interior finishing, this is not the case. Traversing halls, the camera is turned into a room to observe the objects of interest. Furthermore, most rooms do not allow low motion with translation (see Fig. 4). Usually a lot of rotation with just a short translation distance is involved. Thus, in many cases, the rotation dominates translation. Scene objects leave the field of view quickly while parallax is low due to the lack of translation. This requirement must be met by the chosen vision-based method. Furthermore, rotating motions of the camera often end up in a similar pose as before when the room is scanned from one position. Hence, loop-closure is advantageous in that case. Considering the abilities of the vision based approaches mentioned above, few alternatives exist that can handle the challenges of a rotation-dominant motion of the camera in indoor scenes. In case during the rotating motion lines that are known from the building model are visible, requirements are not so strict, but this cannot be assumed all the time. The worst case indeed derives from the scenario that no external knowledge about the visible structure is given and camera motion has to be estimated totally from sensor measurements. Thus, VO and key frame based VSLAM methods may fail under these conditions, but the filter-based VLSAM approach appears to be the right choice if robustness needs to be preferred.

**Far and close observation range:** Indoor scenes provide, on the one hand, narrow corridors and small rooms, but, on the other hand, can show extensive halls. Since the field of view of the camera system is not constrained, it cannot be assumed that a part of the observed view is a close field. The vision methods need to handle the environment no matter which distance is to be expected. Here, binocular camera systems and RGB-D cannot give a precise depth estimation, if at all, since binocular systems degenerate to a monocular system if the baseline relative to the observed distance is too low. Also, RGB-D sensors have a very limited range. Current methods working with this depth information rely on this fact and are very dependent on it.

**Uncontrolled/unconstrained motion:** In the simplest case, the camera system is traversed through the building by personnel at the construction site. The personnel that is assigned for this task might be trained with respect to limitations of a vision based approach, but certainly cannot mimic the steady motion of, for example, a robot. Thus, the vision-based method needs to be aware of unstable motion. Furthermore, the motion cannot be constrained in order to reduce degrees of freedom or enable absolute pose estimation, i.e. with respect to the ground plane. This means, that constrained approaches that reduce the degree of freedom cannot be assumed. All vision-based methods provide approaches that can handle such motion.

**Structured environment and low textured views:** Low amount of features can be extracted in some cases when only a minimally textured scene is observed (see Fig. 4a). This may present some methods with problems when a large amount of features are necessary for computation. Furthermore, features must be observed over a certain distance and thus over a sequence of images. The chosen method must be robust enough to avoid collapsing on
these issues.

Vision-based approaches that allow for a very low amount of features to continue robust motion estimation of the camera have an advantage regarding this aspect. In this constellation, filter-based VSLAM methods are very promising, since even a very little amount of re-measured features can update the filter. With a constant velocity motion model assumption, filter-based approaches work well and can initialize features with new landmarks and in the right scale. SFM confronted with this aspect appears to be inappropriate when orientation-invariant features are used and matching is not constrained in terms of location, since many features can inherit similar descriptors.

**Insufficient light conditions:** Indoor characteristics do not provide perfect conditions for visual sensors, not least due to prevailing light conditions where artificial light sources are considered inadequate during most of the interior finishing phase. Also, this intensifies the noise on the sensor affecting feature extraction. Furthermore, missing inner light sources may intensify shadows from incoming light through openings like windows (see Fig. 4b). Thus, the range of light intensities in images is high from some perspectives and auto-adjustment of camera settings like aperture or exposure may vary. Features can easily be misinterpreted or classified as non-matches. Therefore, it is important to recognize landmarks after they were discarded during the matching process in one or more images. Considering this aspect, vision-based methods that allow for a redetection of images are favorable. Filter-based VSLAM methods are an option that meet these requirements. As well, SFM methods are able to detect matches across several non-matched frames by matching all frames against each other. Since no active matching is available and thus not constrained on matching, SFM might classify more outliers than necessary and sometimes create an interrelation between images.

<table>
<thead>
<tr>
<th>Table 1. Vision-based methods matching indoor requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motion within BIM</strong></td>
</tr>
<tr>
<td>Data of building model not visible</td>
</tr>
<tr>
<td>Rotation-dominating motion supported</td>
</tr>
<tr>
<td>Far and close observation range</td>
</tr>
<tr>
<td>Uncontrolled/unconstrained motion</td>
</tr>
<tr>
<td>Structured environment and low textured views</td>
</tr>
<tr>
<td>Insufficient light conditions</td>
</tr>
</tbody>
</table>

Figure 5. Vision-based method with building model input and image data reveal registered motion information

The comparison of vision-based methods with the necessary requirements of an indoor application during the construction phase are summarized and rated in Table 1. It turns out that VSLAM methods meet the all requirements for this purpose. In particular, an approach that can process line segments with a filter based VSLAM method, and integrate a geometric model and acquire new measurements from the image sensor fulfills these demands. Therefore, this method is accepted as a foundation for a concept that is presented in Fig. 5. This concept includes both image data from a camera sensor and the current building model provided by 4D BIM. The building model provides a set of lines that can initially be loaded into the map of the filter-based VSLAM method. These lines can then be recognized in the images as further measurements that update the motion of the system and the landmarks. Hence, measurements can correct the as-planned lines that could be inaccurate in terms geometry compared to the as-performed positions in 3D space. In order to achieve a robust system and to prevent the lack of features in different poses, additional non-modelled lines are loaded into the map as new landmarks. This has
the advantage that the additional lines that receive their geometric description solely out of the image measurements can already be defined correlating to already well-defined lines out of the building model. This whole process reveals the registered motion information that was targeted by this paper.

5. PRELIMINARY RESULTS

According to the proposed concept in the previous section, the realization of the approach is briefly addressed and preliminary results are illustrated. The first issue that needs to be solved in terms of realization is the acquisition of the first line set. From the correspondences between the line segments from the initial image and the 3D building model lines, the absolute pose can be derived. Since this already requires a coarse approximate pose, this step is performed assisted and thus semi-automatically for the first initial frame. This means, that a pose is first set manually in a 3D viewer that shows the building model superimposed onto the image (see Fig. 6a). The view then roughly shows the desired scene. By applying a robust line registration algorithm, outlier candidates are detected and eliminated from the set of correspondences. Simultaneously, the absolute pose of the image frame within the building model is released (see Fig. 6b).

Extracted lines from images unfortunately rarely have the expected length according to the 3D line that was projected to the image plane. On the one hand, the recognition of lines is difficult even for image-to-image tracking. On the other hand, the initial transfer of line correspondences cannot be performed by simply integrating the 3D lines into the VSLAM system. This presents the whole initialization with problems. The 3D line segments need to be adopted to the image line segments before using them in the VSLAM approach. Currently, this is solved by a method that determines the least distance of the 3D lines according to their representation on the image plane compared to the detected line correspondences in the image (see Fig 6c).

Lines that are to be re-visited by the VSLAM system need to be recognized. Therefore, a line descriptor is applied on the line segments and several further constraints are used. The line descriptor examines the environment of the lines, which can reveal an extra degree of similarity between lines.

The realization of the proposed method is advanced and will be tested soon. The tests will be performed in terms of robustness of the chosen approach for pose estimation in indoor environments. Thereby, a number of aspects needs to be regarded that can influence the accuracy of this approach, including the sensitivity. Since the developed method will be integrated into the context of progress monitoring, the following steps, such as task and object identification in the automation pipeline, will be applied. As well, the chosen approach will be validated for sufficiency with respect to the succeeding steps.

![Figure 6. Illustration of the preliminary implementation results: a) lines present in the image view matched to building model lines and registered to BIM, b) estimated pose of the camera within the building model, c) the representation of the line segments as landmarks within the VSLAM map re-visited after a number of frames](image)

6. CONCLUSIONS & OUTLOOK

The objective of the paper is the investigation of current vision-based methods in terms of their applicability for indoor progress monitoring purposes. Thus, in this paper, vision-based methods that estimate the motion of a camera in a scene are discussed in detail to give an overview with respect to the abilities, strengths and limitations. Afterwards, challenges and requirements that vision-based methods must meet are investigated considering indoor characteristics. A concept is found that fulfills all stated requirements on indoor localization: Using a filter-based VSLAM method integrating BIM knowledge in form of lines that are extracted out of the building model in order to provide registration in terms of scale and pose.

Currently, the concept is being implemented, validated and tested in terms of accuracy compared to ground truth data. Furthermore, the results are matched against results of other vision-based concepts. This concept and the resulting output of registered motion information with BIM will be the base for further sophisticated methods that will allow for a higher degree of automation in indoor progress monitoring.