Multimodal 2D Image to 3D Model Registration via a Mutual Alignment of Sparse and Dense Visual Features

Nathan Crombez, Ralph Seulin, Olivier Morel, David Fofi, Cédric Demonceaux

Multimodal 2D Image to 3D Model Registration via a Mutual Alignment of Sparse and Dense Visual Features

Nathan Crombez, Ralph Seulin, Olivier Morel, David Fofi and Cédric Demonceaux

***Abstract*—Many fields of application could benefit from an accurate registration of measurements of different modalities over a known 3D model. However, aligning a 2D image to a 3D model is a challenging task and is even more complex when the two have a different modality. Most of the 2D/3D registration methods are based on either geometric or dense visual features. Both have their own advantages and their own drawbacks. We propose, in this paper, to mutually exploit the advantages of one feature type to reduce the drawbacks of the other one. For this, an hybrid registration framework has been designed to mutually align geometrical and dense visual features in order to obtain an accurate final 2D/3D alignment. We evaluate and compare the proposed registration method on real data acquired by a robot equipped with several visual sensors. The results highlights the robustness of the method and its ability to produce wide convergence domain and a high registration accuracy.**

I. INTRODUCTION

Nowadays, it becomes relatively simple to create a 3D virtual representation of a real environment. Indeed, vision-based 3D reconstruction methods like SLAM (Simultaneous Localisation and Mapping) [1], SfM (Structure from Motion) [2] or MultiView-Stereo (MVS) [3] are more and more mature. In parallel, technological advances have enabled the development of tools like terrestrial laser scanners (TLS). These methods and devices can now be considered as *out-of-the-box* solutions to create a 3D model of a real scene.

After an environment has been digitized, it may be interesting and useful to supplement the 3D model with novel measurements that come from different visual sensors. However, the use of a diversity of sensors is a source of issues generally grouped under the term "multimodality". Registering a 3D model over 2D images is already a challenging task and is even more complex when the two have been obtained with different visual sensors. Of course, the types of modality that are used depend on the purpose of the application. For instance, near-infrared (NIR) spectral images are commonly used for precision agriculture applications. In [4] authors have designed and developed a multi-spectral 3D imaging device that can be used for creating a 3D point cloud of a field. In addition to geometric and photometric information, each 3D point of the resulting model has also a NDVI (Normalized Difference Vegetation Index) value which is an important indicator of plant vigor. Combining long-wavelength infrared (LWIR) images with a photometric 3D model can reveal information which may not be present neither in the model or in the LWIR images. This facilitates the visual detection, recognition and segmentation of objects like windows on building façades [5]. Fusion of thermal radiation and a 3D model is also useful for monitoring energetic performances of buildings [6] or to study the thermal properties of materials [7]. In cultural heritage documentation, registration of visual data acquired from various 2D and 3D sensing modalities is also a crucial point for the visualisation of big multimodal data [8] or for photorealistic modeling [9]. Finally, registration of 2D image to 3D range scans collected in urban scenarios serves as a core module in many applications [10]. All these works are some examples among many others that show the importance of an accurate multimodal 2D/3D registration.

Authors are from the LE2I laboratory (Laboratoire d'Electronique, Informatique et Image), University of Burgundy, 12 Rue de la Fonderie, 71200 Le Creusot, France email : firstname.lastname@u-bourgogne.fr

II. RELATED WORK

The registration of a 3D model over an image can be seen as the alignment of visual correspondences extracted from these two data. These visual correspondences are generally referred as "visual features". As for classical registration (as opposed to the notion of multimodality), the state-of-the-art approaches for multimodal 2D/3D registration may be broadly classified according to the type of used visual features. They can be sparse or dense and both have advantages and drawbacks.

Sparse feature-based registration requires the extraction and matching of corresponding visual features in the real image and in a virtual image rendered within the 3D model. The geometrical features that are the most commonly used are interest points. For instance, [11] proposed a robust approach for detecting reliable feature correspondences between an image acquired with a range camera and a thermal image by exploiting wavelength independent properties. An EPnP (Efficient Perspective-n-Point) algorithm is used on the resulting set of 2D/3D correspondences to estimate the parameters of the thermal camera and perform the thermal mapping on the 3D data. Other geometrical features like lines have also been used. For instance, contours-based 2D/3D registration method has been used in [12] to align historical painting over 3D model obtained from current images of a scene. After an initial coarse alignment using a shape descriptor, oriented edge points are extracted and matched from contours that are detected in the historical painting and in 3D model renders. Finally, an ICP-like approach is applied on these matched edges to perform the registration. Dominant lines are often preferred in the case of man-made environments as in [13]. Geometrical sparse features require an accurate detection in images of different modalities. They also have to be correctly matched together and even tracked for some approaches. Even if features detection, matching and tracking have been deeply

studied, they are still hard challenges but are crucial to the success of sparse feature-based 2D/3D registration.

As opposed to sparse features, the second category of 2D/3D registration methods are based on dense features. Dense features concept is based on the global appearance of a scene instead of its geometrical shape. A dense feature uses all image pixels. The most commonly used for multimodal 2D/3D registration is the Mutual Information (MI). This statistical measure of non-linear correlation between two data sources was first introduced for the registration of multimodal medical images. The use of MI has been extended to 2D/3D multimodal registration in [14]. Indeed, authors proposed to estimate the camera parameters by maximizing the correlation between a real image and different attributes of illumination of the 3D model (ambient occlusion, specularity, normal field). Very recently, [10] proposed to only use similarity measurements between a chosen set of 2D/3D attribute-pairs that could be dominant in a specific scene. The choice of the attributes-pairs results of a preliminary training phase. Finally, all the selected attributes-pairs are combined into a reliable similarity measurement: Normalized Mutual Information (NMI). NMI have also been used for autonomous vehicles localization based on a LIDAR map of urban environments [15], [16]. Dense features have the advantages to avoid the detection, matching or tracking stages and offers a very accurate registration [14]. However, the convergence domain is very tight, thus to guarantee a correct 2D/3D registration, the real image and the 3D model must be initially coarsely aligned.

These two types of visual features are typically used in two consecutive stages. First, sparse features are used to estimate a first coarse alignment, then a dense feature is used to obtain a fine registration. However, the accuracy of such approaches is highly related to the success of the first phase. An hybrid method has also been studied in [17]. Authors introduced Mutual Correspondences (MC), a semi-automatic 2D/3D registration method based on a minimization function that combines sparse correspondences and MI measure. MC is defined as a simple weighted sum of the two.

In this paper, we propose a robust method that perform accurate multimodal 2D/3D registration that is comparable to an hybrid approach. We did not try to develop a new hybrid similarity measure but we take advantages of both geometrical and dense visual features strengths in an elegant framework. The proposed framework has been designed so that geometrical and dense visual features mutually improve the registration in order to perform a correct and accurate final 2D/3D alignment. Thanks to the use of both feature types, the method has wide convergence domain and a high registration accuracy, regardless of the quality of both the image and the 3D model. Even if the method has been developed in order to perform the alignment of multimodal data, it remains generic and also usable in a classical 2D/3D registration. Moreover, the proposed approach is not limited to a specific pair of geometrical/dense feature types.

This paper is organized as follows. Section III states the problem and describes the several stages of the proposed registration method. Then, experimental results, including qualitative results and quantitative evaluation are presented in Section IV. Finally, conclusions are given in Section V.

III. PROPOSED FRAMEWORK

The registration of a real 2D image over a virtual 3D model can be seen as an estimation of the parameters (intrinsic and extrinsic) of the real camera. This is commonly formalized as an optimization problem. The optimization techniques that are generally used are based on the gradient or the Hessian of the cost function such as Gauss-Newton, Levenberg-Marquardt or Barzilai-Borwein methods. In this work, we propose to use a PSO (Particle Swarm Optimization) approach to perform the registration. PSO solves a problem by having a population (swarm) of candidate solutions (particles) that move around in the search-space. Each particle displacement is influenced by the best particle in its nearest neighborhood and by the best particle in the complete search-space. The particle velocities updated in this way are expected to iteratively move the swarm toward the best solution. PSO is well suited to solve 2D/3D registration problem. First, an analytical derivative of the cost function w.r.t. the camera parameters is not required. Second, having a several virtual cameras appears to be beneficial for both geometrical and dense visual features (Section III-C).

A. Problem formulation

The process inputs are a 3D model noted $\mathbf{O}$ (for *Object*) of an environment and a real image $\mathbf{I}_d$ that may have a different modality. The 3D model is composed of points $\mathbf{P}$ that have 3D coordinates and an intensity value I. The coordinate of a point expressed in the 3D model is noted ${}^o\mathbf{P} = [{}^oX, {}^oY, {}^oZ, 1]^\top$. A 3D model is then a list of N points and intensities: $\mathbf{O} = [{}^o\mathbf{P}_1, I_1], [{}^o\mathbf{P}_2, I_2], ..., [{}^o\mathbf{P}_N, I_N]$. The pose of a virtual camera C (for *Camera*) is described by the following homogeneous transformation matrix ${}^c\mathbf{M}_o$:

$$ {}^c\mathbf{M}_{o(4\times4)} = \begin{pmatrix} & {}^c\mathbf{R}_{o(3\times3)} & & {}^c\mathbf{t}_{o(3\times1)} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1} $$

where ${}^c\mathbf{R}_o$ is a rotation matrix and ${}^c\mathbf{t}_o$ is a translation vector. In the following, we also expressed a camera pose by a 6-element vector ${}^c\mathbf{r}_o = [{}^ct_{Xo}, {}^ct_{Yo}, {}^ct_{Zo}, {}^c\theta_{Xo}, {}^c\theta_{Yo}, {}^c\theta_{Zo}]$ where ${}^c\mathbf{t}_o = [{}^ct_{Xo}, {}^ct_{Yo}, {}^ct_{Zo}]$ and $[{}^c\theta_{Xo}, {}^c\theta_{Yo}, {}^c\theta_{Zo}]$ are Euler angles. The velocity of camera c expressed relatively to the model is noted ${}^c\mathbf{v}_o = {}^c\dot{\mathbf{r}}_o$.

We express a virtual image rendered by the virtual camera C by:

$$ \mathbf{I} = \mathbf{K}pr({}^\mathbf{c}\mathbf{M}_\mathbf{o}\mathbf{O}) \tag{2} $$

where the matrix $\mathbf{K}$ contains the intrinsic parameters (focal length in terms of pixels, principal point and distortion parameters) and the operation $pr(.)$ denotes the projection model of the camera (e.g. perspective, fisheye, omnidirectional...).

The aim of a 2D/3D registration is to find the optimal virtual camera pose that maximizes the similarity, or minimize the difference, between $\mathbf{I}_c$ and $\mathbf{I}_d$, which can be expressed as:

$$ {}^c\hat{\mathbf{M}}_o = \arg\max_{{}^c\mathbf{M}_o} \left[S(\mathbf{f}_c, \mathbf{f}_d) \right] \tag{3} $$

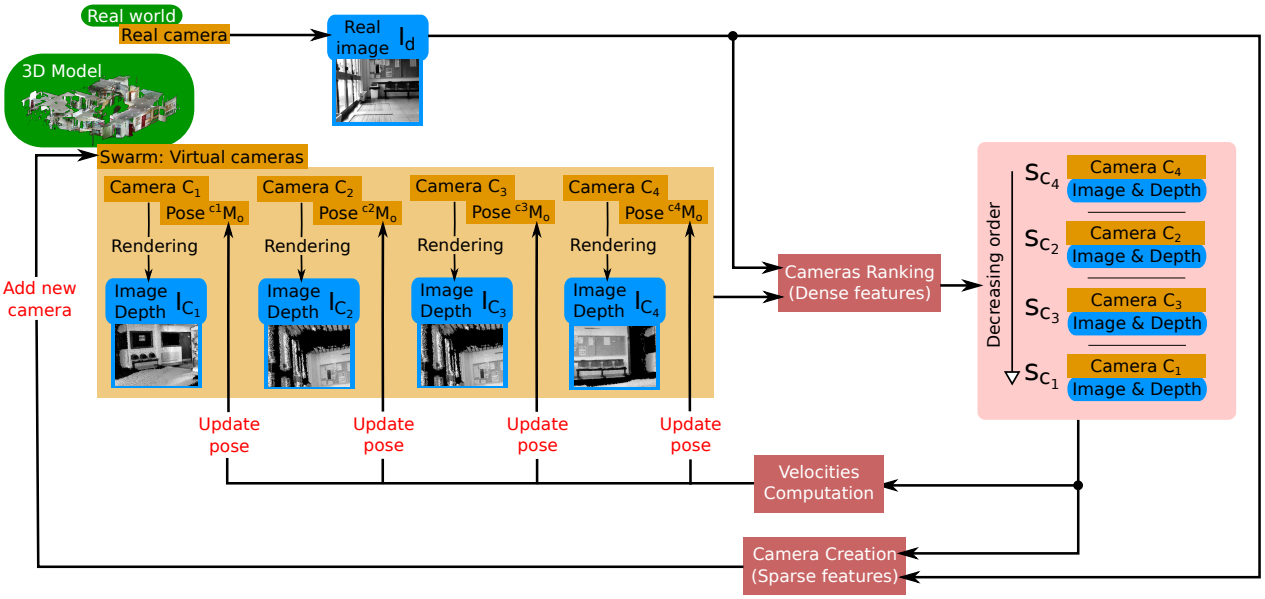


Fig. 1: General overview of the 2D image to 3D model registration method. The stages colored in red represent the core of the approach. For ease of reading, the swarm contains only 4 virtual cameras.

where $\mathbf{f}_c$ and $\mathbf{f}_d$ describes the features extracted from respectively $\mathbf{I}_c$ and $\mathbf{I}_d$ and $S(.)$ is a similarity metric (in the case of a maximization).

A general overview of the proposed framework is illustrated in Fig. 1. The method consists in a swarm of N virtual cameras C_i for $i = [0,...,N]$ that move inside the 3D model $\mathbf{O}$ trying to reach a desired pose represented by the real image $\mathbf{I}_d$. Each virtual camera, in other words each particle, is defined by :

$$C_i \begin{cases} \mathbf{I}_{c_i} & : \textit{Image} \\ {}^{c_i}\mathbf{M}_o & : \textit{Pose} \\ \mathbf{K}_{c_i} & : \textit{Intrinsic Parameters} \\ {}^{c_i}\mathbf{v}_o & : \textit{Velocity} \\ s_{c_i} & : \textit{Similarity Score} \end{cases}$$

The main parts of the proposed registration method are described in the following.

B. Initialization of the swarm

To be able to correctly register the 3D model over the real image, the pose of the real camera has to be inside the search-space of the swarm. Every camera pose ${}^{c_i}\mathbf{M}_o$ is initialized from a random position and a random orientation. The position of the N virtual cameras in the 3D model are randomly initialized inside a sphere around a specific position. Orientations are randomly initialized within a cone. The sphere radius and the angle ranges that defined the orientation cone represent the limits of the search-space. The velocity of every virtual camera is zero in the initial state of the method.

If the real camera has been calibrated, its intrinsic parameters $\mathbf{K}_d$ are already known and all the virtual cameras can be configured with it $\mathbf{K}_{c_i} = \mathbf{K}_d \ \forall \ i = [0,...,N]$. Otherwise the intrinsics parameters of the virtual cameras have to be initialized randomly around a reasonably range and will be estimated in addition to the pose. For ease of reading in the following, we consider that the real camera has been calibrated, thus only the pose is optimized.

At this point, we have N virtual cameras intrinsically and extrinsically initialized and N virtual images rendered from every camera pose. The next stages of the method represent the core of the registration.

C. Evolution of the swarm

The core of the registration process consists of two phases: virtual cameras displacement and virtual camera creation. Each phase is based on a different type of visual features.

1) Displacement phase: The displacement of the N virtual cameras is based on dense visual features. Thanks to the use of every pixel of the images, dense features have the advantage to provide a global minimum clearly defined leading to a very accurate registration. However, it is well known that their cost function are highly non-linear. Having many virtual cameras theoretically reduces the probability of being trapped in a local minimum during the registration process. Considering the different modalities between the 3D model (by extension, the virtual images) and the real image, the state-of-the-art has shown that the Mutual Information (MI) and its derivatives are well suited as similarity metrics [10][14]. In the case of MI as dense metric, equation (3) then becomes:

$${}^{c}\hat{\mathbf{M}}_o = \underset{{}^{c_i}\mathbf{M}_o}{\arg\max} \ [MI(\mathbf{I}_d, \mathbf{I}_{c_i})] \ \forall \ i = [1,...,N] \tag{4}$$

with

$$MI(\mathbf{I}_d, \mathbf{I}_{c_i}) = H(\mathbf{I}_d) + H(\mathbf{I}_{c_i}) - H(\mathbf{I}_d, \mathbf{I}_{c_i}) = s_{c_i} \tag{5}$$

where $H(\mathbf{I}_d)$ and $H(\mathbf{I}_c)$ are individual entropies and $H(\mathbf{I}_d, \mathbf{I}_c)$ is the joint entropy.

We assume that the nearer a virtual camera is to the real camera, the more its rendered image is similar to the real one. Consequently we consider that the more a virtual camera has a high similarity score the better is its pose in the search-space. At each iteration of the PSO algorithm, the N virtual cameras move in the direction of the camera which has the highest similarity score. Their movements are also influenced by the best particle in their nearest neighborhood. More precisely, velocities are updated at each iteration following:

$$\begin{aligned} {}^{c_i}\mathbf{v}_o(t+1) = & \gamma\ {}^{c_i}\mathbf{v}_o(t) + \mu_1\omega_1({}^{c_g}\mathbf{r}_o(t) - {}^{c_i}\mathbf{r}_o(t)) \\ & + \mu_2\omega_2({}^{c_l}\mathbf{r}_o(t) - {}^{c_i}\mathbf{r}_o(t)) \end{aligned} \tag{6}$$

where γ is an inertia factor, μ_1 and μ_2 are acceleration constants and ω_1 and ω_2 are random weight that are distributed uniformly in $[0,1]$. C_g denotes the best camera of the complete swarm and C_l denotes the best camera in the local neighborhood of each camera. The configuration of the PSO parameters is discussed in the experiments section (Section IV-B).

The pose of every virtual camera in the 3D model is updated following:

$${}^{c_i}\mathbf{M}_o(t+1) = {}^{c_i}\mathbf{M}_o(t)\ e^{[{}^{c_i}\mathbf{v}_o(t+1)]} \tag{7}$$

where $e[.]$ is an exponential map of special Euclidean group $SE(3)$ used to determine a displacement from a velocity vector. The velocities of the cameras are updated iteratively until a stop criterion is reached (e.g. a maximum number of iterations or a threshold on the spatial distribution of the cameras). At the end of the process, all the cameras are supposed to have iteratively converged to the best solution ${}^{c}\tilde{\mathbf{M}}_o$ in order to solve equation (4). Even if PSO has a high global search capacity, considering the high non-linearity of our cost function (equation (3)) we have no guarantee to find the optimal solution. To overcome this weakness, just before the computation of the cameras velocity (Fig. 1), a second phase implementing sparse features, tries to create a new virtual camera to add to the swarm.

2) Creation phase: having many virtual viewpoints of the scene increases the chance of finding correspondences between images. This is particularly interesting since it is already difficult to accurately detect and match visual feature between real and virtual images and it is even more challenging when these images have different modalities. Very recently, points matching techniques between multimodal pairs of images (visible, thermal, TLS intensity and range images) have been evaluated [18]. Authors have shown that good results are obtained when point features are detected using MSD (Maximal Self-Dissimilarity) and described with SIFT (Scale-invariant feature transform). Based on this study, our creation phase exploit this combination of detector and descriptor.

The N virtual cameras are reordered in a decreasing order regarding to the visual dense similarity scores computed in the previous phase: $s_{c_i} > s_{c_{i+1}}\ \forall\ i = [1,...,N]$. Point features are matched between the real image $\mathbf{I}_d$ and the best virtual image $\mathbf{I}_{c_1}$ that gives us a list of theoretical 2D/2D correspondences:

$${}^{d}\mathbf{u}^{c_1} \longleftrightarrow {}^{c_1}\mathbf{u} \tag{8}$$

$${}^{d}\mathbf{x}^{c_1} \longleftrightarrow {}^{c_1}\mathbf{x} \tag{9}$$

where ${}^{d}\mathbf{u}^{c_1}$ and ${}^{c_1}\mathbf{u}$ are expressed in the image space and where ${}^{d}\mathbf{x}^{c_1}$ and ${}^{c_1}\mathbf{x}$ are expressed in the normalized metric space following respectively: ${}^{d}\mathbf{x}^{c_1} = \mathbf{K}_d^{-1}\ {}^{d}\mathbf{u}$ and ${}^{c_1}\mathbf{x} = \mathbf{K}_{c_1}^{-1}\ {}^{c_1}\mathbf{u}$. The 2D points ${}^{c_1}\mathbf{x}$ are back-projected in 3D using: ${}^{c_1}\mathbf{X} = pr^{-1}({}^{c_1}\mathbf{x})$. The resulting 3D points are then expressed in the frame of the model $\mathbf{O}$ with: ${}^{o}\mathbf{X}^{c_1} = {}^{o}\mathbf{M}_{c_1}\ {}^{c_1}\mathbf{X}$. This leads to a set of 2D/3D correspondences:

$${}^{d}\mathbf{x}_{c_1} \longleftrightarrow {}^{o}\mathbf{X}^{c_1} \tag{10}$$

Knowing these theoretical 2D/3D correspondences, the pose computation of the real camera consists in solving the well known PnP problem:

$${}^{d}\mathbf{x}_{c_1} = pr({}^{c}\tilde{\mathbf{M}}_o\ {}^{o}\mathbf{X}^{c_1}) \tag{11}$$

Due to the differences in terms of the type, visual aspect, modality and due to the 3D model accuracy, it is challenging to find reliable correspondences between the real and the virtual images. Consequently some matches among the set of correspondences (equation (10)) are wrong. For this reason we solved the PnP problem with a RANSAC (Random Sample Consensus) approach. Indeed, RANSAC uses the smallest set of potential correspondences (4 matches in our case) and iteratively tries to expand this set with consistent data. The algorithm provides the camera pose ${}^{c}\tilde{\mathbf{M}}_o$ that solves equation (11) and also classifies the 2D/3D matches (equation (10)) as inliers and outliers.

If the inliers are good enough, the estimated camera parameters ${}^{c}\tilde{\mathbf{M}}_o$ should be inside the initial search-space (Section III-B). In this case a new virtual camera that is added to the swarm with ${}^{c}\tilde{\mathbf{M}}_o$ as initial pose and we progress to the next step (Figure 1). If ${}^{c}\tilde{\mathbf{M}}_o$ is outside the initial search-space, we extract and match point features between the real image and the next best image $\mathbf{I}_{c_2}$ of the swarm. The resulting 2D/3D correspondences are added to the inliers of the previous points matching:

$$\left\{inlers({}^{d}\mathbf{x}_{c_1}), {}^{d}\mathbf{x}_{c_2}\right\} \longleftrightarrow \{inliers({}^{o}\mathbf{X}^{c_1}),\ {}^{o}\mathbf{X}^{c_2}\} \tag{12}$$

This new set of correspondences is used as before to solve the PnP problem (equation (11)). This procedure is repeated until the estimated ${}^{c}\tilde{\mathbf{M}}_o$ falls within the initial search-space. If ${}^{c}\tilde{\mathbf{M}}_o$ is outside the search-space even using every virtual images, no camera is added during the current creation phase. However, at the next iteration, all the cameras will have moved towards the current best one (highest dense score), consequently the virtual images will be different and the next creation phase has more chance to lead to a better set of correspondences.

Adding a new virtual camera to the swarm is actually always beneficial. If the new camera is actually well estimated, it may be the future global best camera of the swarm. The others will then move in its direction and the dense feature-based PSO should locally improve this solution. If the new camera

is estimated inside the search-space but is not close to the desired solution, it will still be a new particle of the swarm that will explore differently the PSO search-space and may be the source of other interesting visual features.

IV. ANALYSIS AND EVALUATION

In a first time, our 2D/3D multimodal acquisition process is detailed. Then, the performances of the proposed approach are qualitatively and quantitatively evaluated.

A. Dataset acquisition

A ground mobile robot (Fig. 2) with several visual sensors has been instrumented to acquire images of different modalities and to simultaneously create a 3D model of the environment. The robotic base is a Summit XL from Robotnik [19] that has been equipped with a Kinect 2.0, a near-infrared camera (AVT Marlin F-131B NIR) and a polarization camera (4D Technology PolarCam). The near-infrared camera with a

Fig. 2: Dataset acquisition - Robotnik Summit XL equipped with a Kinect 2.0, a near-infrared camera and a polarization camera

$2/3$" sensor and a 8mm f1.4 lens has a narrow field of view. The polarization camera with a $1/2$" sensor and a 1.4mm lens has super wide angle up to $185°$. The three cameras have been calibrated. Thus, their intrinsic parameters and the relative transformations between each sensor are known.

The complete system architecture including camera interfaces, localization and mapping is ROS (Robot Operating System) [20]. The visual SLAM ORB-SLAM2 [1] is applied on the Kinect data to compute the robot trajectory. The keyframes selected by the ORB-SLAM2 algorithm are very well localized thanks to local bundle adjustment, loop closure detection and pose graph optimization. Knowing the pose of the Kinect (and by extension the pose of the two other cameras) at every keyframe, we merge every 3D point clouds acquired at every keyframe in order to build a dense 3D reconstruction with true scale of the environment. Fig. 3 shows a 3D model created following this approach on 470 keyframes. The robot did a loop trajectory of 47.2 meters to create a 3D model that covers an area of about 16×24 meters.

In addition to the Kinect data, the near-infrared and the polarization images and their corresponding camera poses are also saved for every keyframes. In summary, the complete dataset contains the 3D model, the 470 images acquired by the three cameras and their poses expressed in the 3D model reference frame. These camera poses serve as ground truth to evaluate the proposed registration method in the next section.

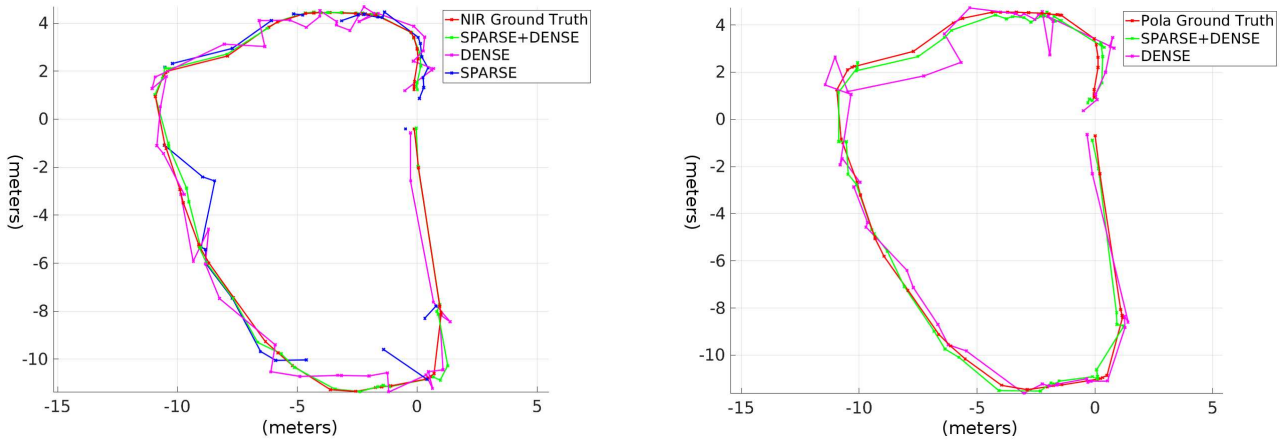
Fig. 3: A 3D model of an indoor building environment created from Kinect data acquired by the Summit XL robot

B. Experimental results

For practical reasons, we have evaluated our method on 50 keyframes randomly selected among the 470 of the dataset. To highlight the advantage of using both visual feature types, we compare the estimation of the poses of the near-infrared and the polarization cameras using the proposed method (DENSE+SPARSE), an approach that uses only sparse features (SPARSE) and another one that only use dense features (DENSE).

The SPARSE approach computes the camera pose using point features (MSD+SIFT) matched between the real image and a virtual image rendered from the Kinect pose. The resulting PnP problem is solved following a standard RANSAC scheme. The DENSE approach estimates the real camera pose by minimizing the mutual information between the real image and virtual images using a PSO. In order to test the convergence properties of our method (DENSE+SPARSE) we deliberately initialize the virtual cameras of the swarm in a very large search-space. The performance of PSO depends on the parameters configuration: number of particles, initialization of the swarm, inertia factor, acceleration constants, stop criteria. Many variants of the PSO algorithm have been proposed to initialize and to optimize the evolution of these parameters. This is not the point of this experiment which is a proof of concept, the parameters are thus chosen empirically. For every keyframe, we initialize a swarm of 30 virtual cameras in a 2 meters diameter sphere centered on the current Kinect position. The cameras orientation is initialized in a range of $\pm 20°$ around the 3 axes. The number of virtual cameras that can be created (Section III-C2) during the registration is limited to 20. The DENSE approach has been initialized with the same parameters.

Fig. 4 shows a comparison between the ground truth and the trajectories of the near-infrared camera and the polarization camera estimated using the 3 methods. As expected, because



(a) (b)

Fig. 4: Comparison of the ground truth and the trajectories of the near-infrared camera (a) and the polarization camera (b) estimated with our method, the sparse features only and the dense features only approaches

of the very large initial search-space and the very tight convergence domain of the method based only on the MI metric, the poses estimated with the DENSE approach are often very far from the desired solution. On the other hand, because of the difficulties to find reliable sparse correspondences between the real and the virtual images, the poses computed with the SPARSE approach are often very badly estimated or even not estimated at all (Fig. 4b). Despite these problems the proposed framework that combines dense and sparse features takes advantages of both and provides an accurate pose estimation. More precisely, TABLE I summarises the mean estimation errors depending on the registration methods.

	Near-Infrared	Polarization
SPARSE	$[51.11cm, 52.47cm, 52.77cm, 3.25°, 4.97°, 3.11°]$	No results
DENSE	$[62.83cm, 92.12cm, 89.76cm, 6.96°, 6.64°, 7.37°]$	$[42.42cm, 70.44cm, 90.05cm, 5.69°, 7.01°, 6.30°]$
SPARSE DENSE	$[6.5cm, 7.3cm, 8.1cm, 0.65°, 0.72°, 0.61°]$	$[15.15cm, 20.86cm, 18.07cm, 3.02°, 5.15°, 4.04°]$

TABLE I: Mean pose estimation errors of the two cameras according to the registration method.

Fig. 5 gives a visual idea about the dissimilarity aspect between the real images and the virtual ones. It also gives a qualitative evaluation of the accuracy of the multimodal 2D/3D registration using the proposed method. Indeed, the two first rows show respectively 4 near-infrared images and the virtual images rendered at the estimated camera poses. Similarly, the two last rows show respectively 4 polarization images and the virtual images rendered at the estimated poses.

It should be kept in mind that the ground truth is directly related to the accuracy of the visual SLAM. The mean errors of estimation have also to be put in perspective with the very high initial search space (± 1 meter along the 3 axes and $\pm 20°$ around them). One can note that we chose MI as similarity metric for the displacement phase but it is not a limit of the method, and other metric could be used and maybe preferable regarding the modalities. Similarly, other matching approaches for the creation phase, maybe more robust to multimodality and image distortions, could be preferred.

V. CONCLUSION

Usual 2D/3D registration approaches are relied either on sparse or on dense visual features. Sparse features offer a high domain of convergence but the resulting alignment is highly related to the reliability of geometric features that have to be extracted and matched between images. Dense-based methods avoid feature detection and matching and provide a very accurate registration but have a very small domain of convergence. Furthermore, the use of a diversity of sensors increases the drawbacks of both types of visual features. In this paper, we have proposed a way to combine sparse and dense features in order to perform automatic and accurate 2D/3D registration. The proposed framework smartly employs both feature types to dramatically increase their strength points. Their combination makes 2D/3D alignment achievable regardless the modalities of both the image and the 3D model. The method has obtained promising results in terms of registration accuracy and robustness even to large initial condition. Of course, the current approach in not free from drawbacks. For further improvements, we plan to study more deeply the influence of the PSO parameters on the quality of the registration.

ACKNOWLEDGMENT

This work is part from a project entitled VIPeR (Polarimetric Vision Applied to Robotics Navigation) funded by the French National Research Agency ANR-15-CE22-0009-VIPeR.

REFERENCES

[1] R. Mur-Artal and J. D. Tardos, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE Transactions on Robotics (TRO)*, vol. PP, no. 99, pp. 1–8, 2017.

Fig. 5: Near-infrared and polarization images given in inputs of the registration method (first and third rows), virtual images rendered at the virtual camera poses estimated using the proposed registration method (second and fourth rows)

[2] C. Wu, "Towards linear-time incremental structure from motion," in *International Conference on 3D Vision (3DV)*, June 2013, pp. 127–134.

[3] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence (PAMI)*, vol. 32, no. 8, pp. 1362–1376, 2010.

[4] J. Das, G. Cross, C. Qu, A. Makineni, P. Tokekar, Y. Mulgaonkar, and V. Kumar, "Devices, systems, and methods for automated monitoring enabling precision agriculture," in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, Aug 2015, pp. 462–469.

[5] D. Iwaszczuk, L. Hoegner, and U. Stilla, "Detection of windows in ir building textures using masked correlation," *Photogrammetric Image Analysis*, pp. 133–146, 2011.

[6] L. Hoegner, S. Tuttas, Y. Xu, K. Eder, and U. Stilla, "Evaluation of methods for coregistration and fusion of rpas-based 3d point clouds and thermal infrared images." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 41, 2016.

[7] A. Bennis, V. Bombardier, P. Thiriet, and D. Brie, "Contours based approach for thermal image and terrestrial point cloud registration," in *XXIV International CIPA Symposium, CIPA 2013*, vol. XL-5/W2, Sep 2013, pp. 97–101.

[8] H. Kim, A. Evans, J. Blat, and A. Hilton, "Multi-modal visual data registration for web-based visualisation in media production," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[9] R. Pintus, E. Gobbetti, M. Callieri, and M. Dellepiane, *Techniques for Seamless Color Registration and Mapping on Dense 3D Models*, ser. Sensing the Past - From artifact to historical site, Nicola Masini AND Francesco Soldovieri (Editors),Sensing the Past, N. Masini, F. Soldovieri (eds.). Springer International Publishing AG, 2017, ch. 17, pp. 355–376.

[10] Y. Zhao, Y. Wang, and Y. Tsai, "2d-image to 3d-range registration in urban environments via scene categorization and combination of similarity measurements," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1866–1872.

[11] M. Weinmann, J. Leitloff, L. Hoegner, B. Jutzi, U. Stilla, and S. Hinz, "Thermal 3d mapping for object detection in dynamic scenes," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 2, no. 1, p. 53, 2014.

[12] B. C. Russell, J. Sivic, J. Ponce, and H. Dessales, "Automatic alignment of paintings and photographs depicting a 3d scene," in *3rd International IEEE Workshop on 3D Representation for Recognition (3dRR-11), associated with ICCV*, November 2011.

[13] L. Liu and I. Stamos, "A systematic approach for 2d-image to 3d-range registration in urban environments," *Computer Vision and Image Understanding*, vol. 116, no. 1, pp. 25 – 37, 2012.

[14] M. Corsini, M. Dellepiane, F. Ponchio, and R. Scopigno, "Image-to-Geometry Registration: a Mutual Information Method exploiting Illumination-related Geometric Properties," *Computer Graphics Forum*, 2009.

[15] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman, "FARLAP: Fast Robust Localisation using Appearance Priors," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, USA, May 2015.

[16] R. W. Wolcott and R. M. Eustice, "Visual localization within lidar maps for automated urban driving," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2014, pp. 176–183.

[17] M. Sottile, M. Dellepiane, P. Cignoni, and R. Scopigno, "Mutual correspondences: an hybrid method for image-to-geometry registration," in *Eurographics Italian Chapter Conference 2010*, Nov 2010, pp. 81–88.

[18] M. Gesto-Diaz, F. Tombari, D. Gonzalez-Aguilera, L. Lopez-Fernandez, and P. Rodriguez-Gonzalvez, "Feature matching evaluation for multimodal correspondence," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 129, pp. 179 – 188, 2017.

[19] Robotnik automation - mobile robot summit xl. [Online]. Available: http://www.robotnik.eu/mobile-robots/summit-xl/

[20] M. Quigley, B. Gerkey, K. Conley, J. Faust, T. Foote, J. Leibs, E. Berger, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, May 2009.